«Uma análise profunda de como podemos proteger a dignidade e os valores humanos numa era de máquinas autónomas.» —WALTER ISAACSON

INTELIGÊNCIA ARTIFICIAL, ESPERANÇA, E O ESPÍRITO HUMANO

GÉNESIS

HENRY A. KISSINGER

CRAIG MUNDIE

ERIC SCHMIDT

PREFÁCIO DE NIALL FERGUSON

Índice

Ficha Técnica

Prefácio

In memoriam: henry a. Kissinger

Introdução

PARTE I – NO COMEÇO

- Capítulo 1 – Descoberta
- Capítulo 2 – O Cérebro
- Capítulo 3 – Realidade

PARTE II – OS QUATRO RAMOS

- Capítulo 4 – Política
- Capítulo 5 – Segurança
- Capítulo 6 – Prosperidade
- Capítulo 7 – Ciência

PARTE III – A ÁRVORE DA VIDA

- Capítulo 8 – Estratégia

Conclusão

Agradecimentos

GÉNESIS

INTELIGÊNCIA ARTIFICIAL, ESPERANÇA, E O ESPÍRITO HUMANO

HENRY A. KISSINGER
CRAIG MUNDIE
ERIC SCHMIDT

COM ELEANOR RUNDE
PREFÁCIO DE NIALL FERGUSON

Tradução de Mário Dias Correia

Ficha Técnica

Título: Génesis
Título original: Genesis
Autores: Henry Kissinger, Eric Schmidt, Craig Mundie
Edição: Duarte Bárbara
Tradução: Mário Dias Correia
Revisão: Vasco Grácio
Capa: Gregg Kulick / © 2024 Hachette Book Group, Inc.
Adaptação de capa: Maria Manuel Lacerda
Imagem de capa: © Getty Images/Flávio Coelho
ISBN: 9789722083805

Publicações Dom Quixote
Uma editora do grupo Leya
Rua Cidade de Córdova, n.º 2
2610-038 Alfragide – Portugal
Tel. (+351) 21 427 22 00
Fax. (+351) 21 427 22 01

©2024, Propriedade de Henry A. Kissinger, Craig J. Mundie e Eric Schmidt ©2024, Publicações Dom Quixote
Prefácio: © Niall Ferguson
Todos os direitos reservados de acordo com a legislação em vigor.
Este livro segue o Acordo Ortográfico de 1990.
www.leya.com

Para o Dr. Kissinger:
homem de Estado, diplomata, mentor e amigo.
Saudamos-te.

PREFÁCIO

Niall Ferguson

Quando, em junho de 2018, Henry Kissinger publicou na revista *Atlantic* o seu ensaio «How the Enlightenment Ends», o facto de o velho estadista ter uma opinião sobre o tema da inteligência artificial surpreendeu muita gente. Kissinger acabava de fazer 95 anos. A IA ainda não era o tema quente que se tornaria depois de a OpenAI ter lançado o ChatGPT no final de 2022.

A mim, como seu biógrafo, pareceu-me perfeitamente natural que o tópico lhe captasse a atenção. Afinal, já em 1957 se tornara conhecido do público com um livro sobre uma tecnologia nova e revolucionária. *Nuclear Weapons and Foreign Policy* era um livro tão minuciosamente pesquisado que mereceu a aprovação inclusive de Robert Oppenheimer, que o descreveu como «extraordinariamente bem informado, e neste aspeto sem precedentes na área do armamento nuclear (...) escrupuloso no seu respeito pelos factos e, ao mesmo tempo, apaixonado e duro na argumentação».

Apesar de ter mergulhado a fundo na história diplomática da Europa do início do século XIX como estudante de doutoramento, ao longo da sua carreira Kissinger teve sempre perfeita consciência de que os padrões eternos da política das grandes potências estavam sujeitos a perturbações periódicas devido às mudanças tecnológicas. Como muitos membros da sua geração que serviram na Segunda Guerra Mundial, viu com os seus próprios olhos não só a morte e a destruição em massa que podiam ser infligidas pelas armas modernas, mas também as terríveis consequências para os judeus, como ele próprio, daquilo a que Churchill chamou memoravelmente a «ciência pervertida» do Terceiro Reich de Hitler.

Desmentindo uma injustificada reputação de «belicista», Kissinger foi, durante toda a sua vida adulta, fortemente motivado pelo imperativo de evitar a Terceira Guerra Mundial – a consequência que todos temiam caso a Guerra Fria entre os Estados Unidos e a União Soviética aquecesse. Sabia muito bem que a tecnologia da fissão nuclear faria de um terceiro conflito mundial um cataclismo ainda mais mortífero do que a guerra contra a Alemanha nazi. No início de *Nuclear Weapons and Foreign Policy*, calcula os efeitos destrutivos de uma bomba de dez megatoneladas lançada sobre Nova Iorque, e em seguida extrapola que um ataque soviético total às cinquenta maiores cidades dos EUA causaria 15 a 20 milhões de mortos e 20 a 25 milhões de feridos. Outros cinco a dez milhões morreriam devido aos efeitos da precipitação radioativa, enquanto talvez outros sete a dez milhões ficariam doentes. Os sobreviventes enfrentariam uma «desintegração social». Mesmo depois de um tal ataque, observa, os Estados Unidos continuariam a ser capazes de infligir uma devastação comparável à União Soviética. A conclusão era óbvia: «A partir de agora, o único resultado de uma guerra total será a derrota de ambos os contendores.» Não poderia haver vencedor num tal conflito, «porque mesmo o lado mais fraco pode ser capaz de infligir um grau de destruição que nenhuma sociedade conseguiria suportar».

Mas o idealismo juvenil de Kissinger não fazia dele um pacifista. No mesmo livro, foi bastante explícito ao afirmar que «os horrores da guerra nuclear não seriam provavelmente evitados por uma redução do armamento nuclear» ou, já agora, por sistemas de inspeção. A questão não era saber se a guerra podia ser evitada por completo, mas saber se era «possível imaginar aplicações de poder menos catastróficas do que uma guerra termonuclear total». Porque se não fosse, então seria sem dúvida muito difícil para os Estados Unidos e os seus aliados prevalecerem na Guerra Fria. «A ausência de quaisquer limites geralmente entendidos para a guerra», avisou, «mina a estrutura psicológica da resistência às intenções dos comunistas. Quando a guerra é considerada equivalente ao suicídio nacional, a rendição pode parecer o menor de dois males.»

Foi neste pressuposto que Kissinger avançou com a sua doutrina da guerra nuclear limitada:

> Contra o sinistro pano de fundo da devastação termonuclear, o objetivo da guerra já não pode ser a vitória militar tal como a entendemos. Em vez disso, deve ser a obtenção de certas condições políticas específicas que o adversário compreenda perfeitamente. O objetivo da guerra limitada é infligir perdas ou colocar o inimigo em risco de uma forma que seja desproporcionada em relação aos objetivos em causa. Quanto mais moderado for o objetivo, menos violenta será a guerra.

Para isso, seria necessário conhecer a psicologia da parte contrária, bem como a dimensão da sua capacidade militar.

Na altura, muitas pessoas recusaram a aparentemente fria aceitação de Kissinger de uma guerra nuclear limitada. Alguns académicos, como Thomas Schelling, contestaram que uma escalada imparável pudesse ser evitada; até o próprio Kissinger se distanciou mais tarde do seu argumento. No entanto, ambas as superpotências continuaram a construir e a utilizar armas nucleares táticas ou de campo de batalha, seguindo precisamente a lógica que Kissinger tinha delineado em *Nuclear Weapons and Foreign Policy*. A guerra nuclear limitada pode não ter funcionado em teoria, mas os planeadores militares de ambos os lados comportaram-se como se pudesse funcionar na prática. (De facto, essas armas continuam hoje a ser fabricadas. O governo russo ameaçou usá-las em mais de uma ocasião desde que a invasão da Ucrânia se revelou um fracasso). O jovem Kissinger tinha mais razão em relação às armas nucleares do que ele próprio imaginava.

Kissinger nunca deixou de refletir sobre as implicações das mudanças tecnológicas no domínio político. Num documento há muito esquecido que escreveu para Nelson Rockefeller em janeiro de 1968, analisava os modos como a informatização poderia ajudar os funcionários a lidar com o fluxo cada vez maior de informações geradas pelas agências governamentais dos EUA. Na sua opinião, os altos funcionários corriam o grave risco de se afogarem em dados. «O decisor político de topo», escreveu, «tem tanta informação ao seu dispor que, em situações de crise, acha impossível lidar com ela.»

Os decisores precisavam de ser «consistentemente informados sobre os prováveis pontos problemáticos», argumentava Kissinger, incluindo os *potenciais* pontos problemáticos «mesmo quando não lhes foi atribuída prioridade máxima». Também precisavam de dispor de «um conjunto de opções de ação (...) que descrevam as principais alternativas em resposta a circunstâncias previsíveis, com uma avaliação das consequências prováveis, ao nível doméstico e internacional, de cada uma dessas alternativas». Kissinger reconhecia que, para conseguir uma cobertura tão abrangente, seriam necessários grandes investimentos em programação, armazenamento, recuperação e gráficos. Felizmente, existia a «tecnologia de *hardware*» para desempenhar estas quatro funções:

> Temos hoje a possibilidade de armazenar várias centenas de itens de informação sobre cada indivíduo nos Estados Unidos numa fita magnética com 730 metros de comprimento. (...) [Os] computadores de terceira geração são capazes de efetuar operações básicas em nanossegundos, ou seja, em bilionésimos de segundo. (...) [E]xperimentalmente, os sistemas de partilha de tempo demonstraram que a capacidade de acesso múltiplo a computadores digitais de grande escala é possível para permitir a entrada/saída de informação tanto nas estações executivas como nas estações operacionais distribuídas por todo o mundo. (...) [E] muito em breve o ecrã de tubo de raios catódicos a cores estará disponível para a saída do computador.

Mais tarde, após o seu primeiro ano na Casa Branca como conselheiro de Segurança Nacional de Richard Nixon, Kissinger tentou obter um computador deste tipo para seu uso pessoal. A CIA negou o pedido, presumivelmente porque Kissinger sem um computador era o máximo que a comunidade da Informação podia suportar.

*

Henry Kissinger nunca se reformou. Nem nunca deixou de se preocupar com o futuro da humanidade. Dificilmente um homem assim ignoraria uma das mais consequentes inovações tecnológicas do seu tempo: o

desenvolvimento e implementação da inteligência artificial generativa. E de facto a tarefa de compreender as implicações desta tecnologia nascente consumiu uma parte significativa dos seus últimos anos.

Génesis, o último livro de Kissinger, foi escrito em coautoria com dois eminentes tecnólogos, Craig Mundie e Eric Schmidt, e tem a marca do otimismo inato desses inovadores. Os autores aguardam com expetativa a «evolução do *Homo technicus* – uma espécie humana que pode, nesta nova era, viver em simbiose com a tecnologia das máquinas». A IA, argumentam, poderá em breve ser aproveitada «para gerar uma nova base de riqueza e bem-estar humanos (...) [que] permitirá pelo menos aliviar, se não eliminar, as tensões de trabalho, classe e conflito que anteriormente dilaceraram a humanidade». A adoção da IA poderá mesmo conduzir a «profundas igualizações (...) entre raça, género, nacionalidade, local de nascimento e antecedentes familiares».

A contribuição do autor mais velho é, no entanto, detetável na série de avisos que constituem o *leitmotif* do livro. «O advento da inteligência artificial é», observam os autores, «uma questão de sobrevivência humana. (...) Uma IA inadequadamente controlada poderia acumular conhecimento de uma maneira destrutiva. (...) As convulsões que em breve irão transformar a realidade coletiva do planeta (...) marcam uma rutura fundamental com o passado.» Aqui, reformulada para o *Génesis* mas imediatamente reconhecível, está a pergunta original de Kissinger no seu ensaio de 2018 publicado na *Atlantic*, «How the Enlightenment Ends»:

> A capacidade objetiva [da IA] de chegar a conclusões novas e precisas sobre o nosso mundo através de métodos não humanos não só perturba a nossa confiança no método científico, tal como tem sido continuamente perseguido desde há cinco séculos, mas também desafia a reivindicação humana de uma compreensão exclusiva ou única da realidade. O que pode isto significar? Será que a era da IA não só não vai impulsionar a humanidade para a frente como, em vez disso, vai catalisar um regresso à aceitação pré-moderna de uma autoridade inexplicável? Em suma: estaremos, poderemos estar, na beira do precipício de uma grande reviravolta na cognição humana – um iluminismo negro?

Naquela que pareceu a este leitor a secção mais poderosa do livro, os autores contemplam uma corrida ao armamento da IA profundamente preocupante. «Se (...) cada sociedade humana quiser maximizar a sua posição unilateral», escrevem os autores, «estarão criadas as condições para uma competição psicológica entre forças militares e agências de informação rivais diferente de tudo o que a humanidade viu até agora. Atualmente, nos anos, meses, semanas e dias que antecedem a chegada da primeira superinteligência, espera-nos um dilema de segurança de natureza existencial.»

Se já estamos a assistir a «uma competição para alcançar uma inteligência única, perfeita e inquestionavelmente dominante», quais são os resultados prováveis? Os autores preveem seis cenários, segundo as minhas contas, nenhum deles aliciante:

1. A humanidade perderá o controlo de uma corrida existencial entre múltiplos atores presos num dilema de segurança.
2. A humanidade sofrerá o exercício de uma hegemonia suprema por parte de um vencedor livre dos tradicionais travões e contrapesos necessários para garantir um mínimo de segurança para os outros.
3. Não haverá apenas uma IA suprema e sim múltiplas instanciações de inteligência superior no mundo.
4. As empresas que detêm e desenvolvem a IA podem vir a acumular um poder social, económico, militar e político totalizante.
5. A IA poderá encontrar a maior relevância e a expressão mais alargada e duradoura não nas estruturas nacionais, mas nas religiosas.
6. A difusão descontrolada e de fonte aberta da nova tecnologia poderia dar origem a bandos ou tribos menores com uma capacidade de IA inferior, mas ainda assim substancial.

Kissinger estava profundamente preocupado com cenários como estes, e o seu esforço para os evitar não terminou com a escrita deste livro. Não é segredo que o último empenho da sua vida – em que esgotou as forças que lhe restavam nos meses seguintes ao seu centésimo aniversário – foi iniciar

um processo de conversações sobre a limitação de armas de IA entre os Estados Unidos e a China, precisamente na esperança de evitar esses resultados distópicos.

A conclusão de *Génesis* é inconfundivelmente kissingeriana:

> À medida que a IA acelerar a linha temporal da evolução para lá do compreensível, a humanidade dividir-se-á em fações que se confrontarão numa luta entre o presente e o futuro. Haverá aqueles que desejam manter a humanidade fixa no seu estado atual, e cuja preferência seria usar a tecnologia para progredir, mas não para se tornarem dependentes dela. Como a linha divisória parece aproximar-se demasiado para alguns, os indivíduos deste grupo poderão recorrer à sabotagem ou possivelmente ao terrorismo, numa tentativa de desencadear um exílio global para um passado mais simples. Outra fação, talvez demasiado confiante nas capacidades da sua criação ou nas suas próprias capacidades, procurará desinibidamente empurrar-nos para um futuro incerto. Seremos engolfados por crises insuperáveis sem uma ajuda superior.

A resposta habitual dos tecnólogos a estes pressentimentos é lembrar-nos os benefícios tangíveis da IA, que já são muito óbvios no domínio da ciência médica. Não discordo. Na minha opinião, o AlphaFold – um modelo baseado em redes neurais que prevê estruturas tridimensionais de proteínas – foi um avanço muito mais importante do que o ChatGPT. No entanto, a ciência médica registou avanços comparáveis no século XX, e mesmo assim as guerras mundiais e o Holocausto aconteceram enquanto eram descobertos e largamente disponibilizados antibióticos, novas vacinas e inúmeras outras terapêuticas.

O problema central do progresso tecnológico manifestou-se durante a vida de Henry Kissinger. A fissão nuclear foi descoberta em Berlim por dois químicos alemães, Otto Hahn e Fritz Strassmann, em 1938. Foi explicada teoricamente (e batizada) pelos físicos austríacos Lise Meitner e o seu sobrinho Otto Robert Frisch em 1939. A possibilidade de uma reação nuclear em cadeia conduzir a uma «produção em grande escala de energia e de elementos radioativos e, infelizmente, talvez também a bombas atómicas» foi ideia do físico húngaro Leó Szilárd. A possibilidade de essa

reação em cadeia poder também ser aproveitada num reator nuclear para gerar calor foi igualmente reconhecida nessa altura. Bastaram, todavia, pouco mais de cinco anos para construir a primeira bomba atómica, ao passo que só em 1951 foi inaugurada a primeira central nuclear.

Pergunte a si mesmo: o que foi que os seres humanos mais construíram nos últimos oitenta anos: ogivas nucleares ou centrais nucleares? Atualmente, há no mundo cerca de 12 500 ogivas nucleares, e o número não para de crescer à medida que a China expande o seu arsenal nuclear. Em contrapartida, existem 436 reatores nucleares em funcionamento. Em termos absolutos, a produção de eletricidade nuclear atingiu o seu pico em 2006, tendo a percentagem da produção total de eletricidade de origem nuclear ao nível mundial diminuído de 15,5% em 1996 para 8,6% em 2022, em parte devido a reações políticas exageradas a um pequeno número de acidentes cujos impactos na saúde humana e no ambiente foram negligenciáveis em comparação com os efeitos das emissões de dióxido de carbono dos combustíveis fósseis.

A lição da vida de Henry Kissinger é clara. Os avanços tecnológicos podem ter tanto consequências benéficas como maléficas, dependendo da forma como decidimos coletivamente explorá-los. É claro que a inteligência artificial é diferente da fissão nuclear numa série de aspetos. Mas seria um erro grave assumir que utilizaremos esta nova tecnologia mais para fins produtivos do que para fins potencialmente destrutivos.

Foi este tipo de perceção, nascida da experiência histórica e pessoal, que inspirou Henry Kissinger a dedicar tanto tempo da sua vida ao estudo da ordem mundial e a evitar a guerra mundial. Foi o que o levou a reagir com tanta alacridade – e preocupação – aos recentes avanços da inteligência artificial. E é por isso que esta publicação póstuma é tão importante como tudo o que ele escreveu ao longo da sua longa e consequente vida.

Oxford,
julho de 2024

IN MEMORIAM:
HENRY A. KISSINGER

O DR. HENRY A. KISSINGER morreu no dia 29 de novembro de 2023. Tinha cem anos. Uma inspiração para todos os que o conheceram, trabalhou até ao fim neste volume, o seu vigésimo segundo livro. Durante as frequentes reuniões que tivemos ao longo do último ano que passou entre nós, insistiu sempre na importância do tema que nos ocupava e na urgente necessidade de difundir a sua mensagem. Nós os dois, como coautores, fomos das últimas pessoas a falar com ele e a vê-lo nos dias que precederam a sua morte. Ao concluirmos agora este projeto, em seu nome e a seu pedido, esforçámo-nos por manter a originalidade do pensamento e o estilo vibrante da voz dessa personalidade notável numa questão da maior importância para o futuro da humanidade. Acabar o que ele começou, para que o último empreendimento literário a que meteu mãos não morra com ele mas viva no mundo sem ele, é apenas uma pequena contribuição para a sua memória.

Depois de ter feito tanto para construir o nosso mundo, consagrou as últimas horas que nele passou a um esforço vital para o salvar. Na realidade, a sua última palavra escrita é um pedido para que todos perseverem no vasto projeto de assegurar o futuro da nossa espécie. Em meados do século passado, o Dr. Kissinger foi um dos principais arquitetos do esforço filosófico e diplomático para proteger a humanidade da aniquilação atómica – o encontro do século XX com as sombrias realidades do risco existencial. O corajoso defensor contra esse perigo parte exatamente no momento em

que um novo perigo se apresenta. A sua vida termina precisamente quando começa uma nova forma de vida. Confrontados com o alvorecer da era da IA, temos consciência de quão poucos além do Dr. Kissinger – estudante do século XIX, mestre do século XX e oráculo do século XXI – alguma vez estiveram tão bem posicionados para lançar as bases do nosso futuro.

O Dr. Kissinger foi, antes e acima de tudo o mais, um filósofo da História. Das suas profundas investigações sobre o tema da tragédia nasceu o projeto a que iria dedicar toda a vida: demonstrar que o idealismo do coração pode ser compatível com o realismo da mente, e até enobrecê-lo. É possível defender, na formulação do escritor francês Romain Rolland, tanto o «pessimismo do intelecto» como o «otimismo da vontade». Onde o otimista aspira a um controlo humano seguro dos assuntos humanos, o pessimista vê a nossa condição como determinada por forças fora do nosso controlo: as leis da natureza e os ciclos da História.

O Dr. Kissinger sabia sem dúvida muito bem como um idealismo fervoroso pode ser aproveitado por ideólogos sem remorsos para derramar sangue ou levar à hesitação no uso do poder. O fascismo, o comunismo, o totalitarismo, o fanatismo religioso militante – todos eles reivindicaram para si os fins mais idealistas perseguidos ao longo da história da humanidade. Vítima desses excessos desumanos, e depois combatente militar e diplomático contra eles, comprometeu-se a ajudar a reconstruir o mundo sobre uma base de ordem sem vergonha e segurança sem culpa. Através da sua gestão ativa dos assuntos internacionais, o Dr. Kissinger conduziu o seu país adotivo – e o mundo – por entre convulsões incertas, mantendo-se firme no sólido terreno dos factos históricos e do interesse nacional.

Brilhante defensor da necessidade seletiva do realismo, o Dr. Kissinger era também um idealista – respeitador, como diz o seu biógrafo Niall Ferguson, do «papel da liberdade, da escolha e da ação humanas na formação do mundo». Na teoria e através da prática, demonstrou a sua convicção de que os seres humanos não vivem, nem podem viver, como se o futuro fosse inevitável. Na sua tese de fim de curso em Harvard, *The Meaning of History*, Henry Kissinger, então com 27 anos, abordava o mesmo debate filosófico que anima agora a sua última obra: «Seja qual for

a nossa conceção da necessidade dos acontecimentos, [e] (...) seja como for que possamos explicar as ações em retrospetiva, a sua realização ocorreu com a convicção interior da escolha.»

Para ele, não era certo que a humanidade sobrevivesse a desígnios desumanos saídos do fogo da sua própria forja. Confrontado com – e sobrecarregado por – a assustadora perspetiva de uma catástrofe nuclear, não sucumbiu ao fatalismo do determinismo nem às profecias de apocalipse. É verdade que os medos existenciais podem dar origem ao niilismo, mas também podem infundir nos melhores de nós a força necessária para desafiar o mal e defender o que precisa de ser preservado para o futuro da nossa espécie. No início dos anos 1950, quando era um jovem professor em Harvard, participou num conjunto de reuniões em que cientistas e académicos de renome como ele se juntaram para discutir e debater as potenciais consequências da guerra nuclear e as medidas necessárias para a evitar. Dessas reuniões nasceram doutrinas que, desde então, têm mantido o nosso mundo a salvo dos piores receios dos participantes.

Décadas mais tarde, conversando connosco, falava frequentemente dessas reuniões – a sua estrutura, o seu objetivo e a sua importância retrospetiva. A sua visão permaneceu igual até ao fim, simultaneamente inconformada com a marcha do destino e refratária a qualquer visão utópica. Ocorre na IA o mesmo equilíbrio que no contexto nuclear: pequenos grupos de indivíduos dedicados podem alterar a história, intervindo e manifestando os seus valores. Ao mesmo tempo, porém, e independentemente do génio dos cientistas que construíam novas inteligências, o Dr. Kissinger acreditava que a formação académica desses cientistas não seria por si só suficiente para garantir um mínimo de segurança no funcionamento dos novos instrumentos.

Por isso o seu legado em matéria de IA não se limita a explorações puramente filosóficas e académicas, engloba também propostas práticas. Meio século depois da sua visita secreta a Pequim e da subsequente abertura das relações entre os Estados Unidos da América e a República Popular da China, o Dr. Kissinger fez mais uma viagem à capital chinesa. Deslocou-se a convite urgente e específico do Presidente Xi Jinping para

discutir, como tema principal, os riscos que a humanidade enfrenta com a IA. Foi a última viagem ao estrangeiro que fez e a sua última missão diplomática.

Se, no passado, o Dr. Kissinger elevou o estudo e a prática da política à condição de forma de arte, hoje a sua procura de respostas fez da IA uma questão de mais do que ciência. Foi, com um de nós e com o Professor Daniel Huttenlocher do MIT, autor de *The Age of AI: And Our Human Future*, publicado em 2021, onde previa que o advento da inteligência artificial ia criar uma nova época na História, semelhante no seu impacto ao Iluminismo do século XVIII pela capacidade de mudar o pensamento humano de maneiras profundas e inesperadas. Nesta nova era, no entanto, em vez de trabalhar a partir de questões levantadas pelos humanos, a humanidade confronta-se com respostas dadas pela IA a perguntas que nenhum humano alguma vez suscitou. Enquanto a IA conquistava os domínios do conhecimento humano, o Dr. Kissinger procurava confiar nos recursos da sabedoria humana e procurar aí as soluções.

Neste volume, sondamos com o Dr. Kissinger o impacto da IA em oito áreas diferentes da atividade e do pensamento humanos, culminando com a sua própria resposta filosófica à incessante procura de uma estratégia que permita equilibrar benefícios e riscos. Na prossecução desse objetivo, o Dr. Kissinger explora as perspetivas de uma coexistência humana com a IA e, a seu tempo, de uma coevolução humano-IA. Ao colocar conceptualmente a possibilidade de uma conciliação destas duas espécies – uma orgânica, a outra sintética – revela também a necessidade de uma escolha: criar um mundo em que a *IA* se torne mais parecida *connosco*, ou um mundo em que *nós* nos tornemos mais parecidos com a *IA*.

Desde a publicação do seu primeiro livro sobre a era da IA, o Dr. Kissinger apercebeu-se cada vez mais da existência de um limite à utilidade da razão nos últimos dias de uma era. Para nós, humanos, as explicações que estejam para lá da nossa compreensão – ou que sejam de nossa própria feitura – podem parecer totalmente desconcertantes; o nosso instinto é presumir que são menos avançadas e mais primitivas do que as nossas explicações científicas, um passo atrás e não à frente. Mas essa é uma suposição perigosa.

Se, nas palavras de Arthur C. Clarke, «Qualquer tecnologia suficientemente avançada é indistinguível da magia», e se os milagres são produtos da matemática, o futuro deveria ser inexplicável, desconcertante, até mágico. Ao longo das décadas em que conhecemos o Dr. Kissinger, ele ensinou-nos generosamente, com base nos seus conhecimentos de política – esse complexo domínio dos assuntos humanos –, que apesar de a razão ter sido o paradigma dominante através do qual os humanos controlaram o nosso mundo, não pode ser o paradigma através do qual nos controlamos a nós mesmos.

Não devemos, pois, à medida que avançamos, ter a esperança de depender apenas da razão – o combustível histórico dos maiores progressos da humanidade. Mas também não devemos abandoná-la totalmente. À semelhança da sua moderação entre idealismo e realismo, as investigações finais do Dr. Kissinger sobre o nosso futuro atingem um equilíbrio entre essa qualidade empírica da verdade e algo mais – filosoficamente para lá da razão, mas cronologicamente anterior a ela. Se a política externa não se pode permitir um excesso de qualquer um dos extremos, o mesmo acontece com o nosso modelo para o futuro.

A IA é, portanto, um desafio único que exige um pensamento que pode parecer, à primeira vista, irracional ou rebuscado – e, de facto, os cenários descritos neste livro são surpreendentes. Mas ao dizer-nos, e aos outros, que ele próprio não passava de um humilde estudante tanto da humanidade como da sua mais recente e potencialmente última criação, instilou em nós a ideia de que o maior perigo posto pela IA seria declararmos demasiado cedo, ou demasiado definitivamente, que a compreendemos.

A sua profundidade de intelecto e a sua perceção das pessoas não são qualidades que esperemos voltar a encontrar. Não conhecemos mais ninguém que, aos 93 anos, tenha conseguido dominar uma área do conhecimento técnico totalmente nova e até então desconhecida. Com a sua insaciável curiosidade e a sua vitalidade mental, aliadas a uma devoção ao trabalho e a um sentido de missão ímpares, nenhuma dor do corpo ou do espírito foi suficiente para afogar a sua paixão pelo progresso. Não obstante as inevitáveis enfermidades da idade avançada, levantava-se todos os dias com a inabalável determinação de empurrar o mundo para a frente. A sua

força indomável vinha, talvez, de uma disciplina sem igual – endurecida na juventude pela opressão, esculpida pelo serviço na guerra e testada durante décadas pelos conflitos da vida pública.

Somos apenas dois dos muitos cujas vidas foram moldadas por este homem extraordinário. Sentiremos muito a sua falta – sem dúvida de mais maneiras do que podemos agora imaginar. Partindo na véspera de uma grande incerteza, ele é agora mais necessário do que nunca. É por isso que, para este livro, parecia não haver um título mais adequado do que *Génesis* – um novo começo para ele e para todos os seres humanos. Quer a humanidade tenha êxito ou fracasse, ele já não estará presente para testemunhar o resultado final dos seus esforços. Pelo menos temos agora a sua sabedoria para guiar os nossos.

Eric Schmidt e Craig Mundie

INTRODUÇÃO

Ainda há bem poucos anos, a inteligência artificial (IA) ocupava um pequeno espaço no debate público. Hoje, na sequência dos rápidos progressos conseguidos, é um tema de primeira página em todos os meios de comunicação social e uma questão que preocupa os líderes da ciência, das empresas, do jornalismo, do serviço público, da educação e da política em todo o mundo.

A nosso ver, tanto o público em geral como muitos especialistas nesta área continuam a ignorar aspetos importantes da nova era da IA. As novas formas de IA e as respostas humanas a elas dadas poderão transformar nada menos do que a relação humana com a realidade e a verdade, a exploração do conhecimento, bem como a evolução física da humanidade, a condução da diplomacia e o sistema internacional. Estas são algumas das questões cruciais das próximas décadas e devem ser as preocupações orientadoras dos líderes em todos os domínios.

As mais recentes capacidades da IA, impressionantes como são, hão de parecer fracas em retrospetiva à medida que os seus poderes aumentam a um ritmo acelerado. Poderes que ainda não imaginámos preparam-se para permear a nossa vida quotidiana. Os futuros sistemas vão facilitar avanços enormes e largamente benéficos, melhorando a nossa saúde e gerando riqueza.

Mas estas capacidades trazem consigo riscos técnicos e humanos, alguns dos quais são conhecidos e outros desconhecidos. As tecnologias atuais já funcionam de modos que os seus inventores não tinham previsto, e é provável que este padrão se mantenha. Cada caminho de investigação

frutuoso seguido pelos nossos cientistas – e eles serão vários – tem a capacidade de dar origem a poderes novos e imprevistos que podem ou não ser compreensíveis ou benéficos para os seres humanos.

A IA parece comprimir as escalas de tempo humanas. Os objetos do futuro estão mais próximos do que imaginamos. Só para dar um exemplo, as máquinas com capacidade para definir os seus próprios objetivos estão quase aí ao virar da esquina. Se quisermos ter alguma esperança de acompanhar os riscos envolvidos, vamos precisar de responder e agir no mais curto espaço de tempo possível. Conscientes do que está em jogo e da urgência da tarefa que temos pela frente, referimos aqui apenas algumas das suas muitas facetas.

À medida que as parcerias homem-máquina se tornarem omnipresentes, os seres humanos terão de determinar a natureza adequada dessas relações. As respostas podem ser tiradas da lógica da segurança e da eficiência, obtidas através do estudo da história, ou percebidas no divino. Indivíduos, nações, culturas e fés vão ter de determinar que limites, se alguns, impor à autoridade da IA sobre a verdade. Vão ter de decidir se permitem que a IA se torne um intermediário entre os humanos e a realidade. Neste contexto, vão ter de escolher entre, por um lado, manter o papel tradicional do empreendimento humano (embora provavelmente cedendo a liderança à IA na descoberta de novos conhecimentos) e, por outro, pôr de lado a mente humana, biologicamente limitada, e estabelecer uma parceria potencialmente reformulada com a IA na fronteira intelectual. Escolhemos os nossos objetivos e utilizamos a IA para os alcançar, ou deixamos que as IA ajudem a escolher sozinhas alguns desses objetivos? Mais urgentemente, a humanidade precisa de dar à dignidade humana uma definição moderna e sustentável que possa fornecer uma orientação filosófica para as nossas decisões nos próximos anos.

O advento da inteligência artificial é, a nosso ver, uma questão de sobrevivência humana. Como explicaremos mais adiante neste livro, as futuras capacidades da IA, funcionando a velocidades inumanas, vão tornar inútil a regulação tradicional. Vamos precisar de uma forma fundamentalmente nova de controlo. Para a comunidade científica global, a tarefa imediata é encontrar medidas técnicas para incluir salvaguardas intrínsecas em todos os sistemas de IA. Por seu lado, as nações e as organizações internacionais, uma vez reunidas em torno de um consenso,

terão de desenvolver – no que às crises diz respeito – novas estruturas políticas de monitorização, implementação e resposta. O que vai exigir a resolução não de um, mas de dois «problemas de alinhamento»: o alinhamento técnico dos valores e intenções humanos com as ações da IA, e o alinhamento diplomático dos humanos uns com os outros.

O Dr. Henry Kissinger, um dos coautores deste livro, orientou de perto os seus dois colaboradores no segundo problema, enquanto eles – como antigos altos funcionários da Microsoft e da Google – o orientaram no primeiro. Craig Mundie foi o principal contacto em questões de política tecnológica da Microsoft com Washington e com governos estrangeiros de todo o mundo, supervisionando simultaneamente a Microsoft Research antes de, mais recentemente, aconselhar as chefias da organização de investigação OpenAI. Eric Schmidt liderou a Google durante uma década antes de passar a década seguinte a criar uma rede de talentos e ideias na interseção de tecnologia, filantropia e segurança com o objetivo de salvaguardar a humanidade.

A urgência das questões com que nos confrontámos em conjunto é tal que, em vez de esperarmos por uma crise, consideramos imperativo que a nossa sociedade e, na realidade, a nossa espécie as abordem de forma proativa. E embora a causa da segurança humana seja uma componente necessária de uma resposta bem-sucedida à IA, não pode responder a todas as questões que ela levanta, porque, na era da IA, a humanidade deixará de ser a mesma. A questão é saber se, e em que medida, nós, os humanos, optaremos por continuar a afirmar a nossa autoridade sobre a maneira como essa mudança ocorre.

PARTE I

NO COMEÇO

CAPÍTULO 1

DESCOBERTA

DESCOBRIR TALVEZ SEJA, tomada isoladamente, a capacidade mais estimulante da espécie humana. Impulsionados pela curiosidade e encantados pela surpresa, preenchemos os vazios que percecionamos e transformamos as perguntas que fazemos em respostas. A exploração é tão essencial para a nossa autodefinição que, apesar dos perigos e frustrações manifestos, continuámos a percorrer incansavelmente os seus muitos caminhos.

Ao longo da história, a exploração humana do ambiente físico, em particular, tem sido uma história de imensa coragem perante perigos extremos. As equipas humanas que se lançaram em tais empreendimentos viram-se muitas vezes confrontadas com a sua própria mortalidade. No início do século XVI, a circum-navegação do globo por Fernão de Magalhães foi uma odisseia de três anos repleta de violência, fome e morte. Magalhães foi o primeiro a traçar uma rota à volta ao mundo. No processo, bateu o recorde de resistência em alto-mar, estabeleceu as dimensões do planeta e – tendo a viagem decorrido no contexto do colonialismo europeu – abriu caminho para o intercâmbio social e económico à escala internacional.

A maior parte dos marinheiros de Magalhães sabia que corria o risco de uma catástrofe. Embora já não se acreditasse que a Terra era plana, a sua rotundidade não tinha ainda sido factualmente provada, e muitos dos membros da tripulação receavam talvez cair da beira do mundo.

Tanto Magalhães como a sua tripulação estavam cientes de que pagariam caro por quaisquer suposições erradas. E de facto erraram, e pagaram: tinham subestimado a duração da viagem, as quantidades de alimentos necessárias para os sustentar, os perigos de envenenamento em massa e o risco de os navios ficarem danificados ou inoperacionais. Da esquadra original de cinco navios com cerca de 270 tripulantes, apenas um navio solitário com dezoito sobreviventes que mais pareciam fantasmas regressou ao porto em Espanha. Entre eles não se encontrava o capitão-mor, que perecera pelo caminho depois de ter sido atingido na perna por uma seta envenenada.[1] Nos 400 anos que se seguiram à época de Magalhães, seriam cartografados todos os recantos do mundo – exceto a Antártida, uma terra tão desolada como um planeta alienígena. O explorador anglo-irlandês Ernest Shackleton iria chegar mais perto do que qualquer outro antes dele de alcançar o Pólo Sul dessa terra: o fim do mundo. Em 1909, com uma tripulação inexperiente e sem qualquer apoio governamental, apenas com alguns empréstimos privados e contribuições individuais, Shackleton e os seus homens estabeleceram o recorde da mais longa viagem deste tipo e abriram caminho aos exploradores que se lhes seguiram.

Apesar de não poder afirmar que foi o primeiro a chegar ao Pólo Sul, Shackleton conquistou o respeito das gerações posteriores. Fê-lo dando prioridade aos valores humanos em detrimento das ambições exploratórias. Um ano após o início da expedição, os membros da equipa, amarrados diariamente a trenós para turnos brutais de dez horas, avançavam apenas alguns quilómetros por dia. Embora tivessem comida suficiente para chegar ao pólo, a que restaria não seria suficiente para lhes permitir regressar ao navio.

Foi então que, a 156 quilómetros do triunfo, Shackleton tomou a decisão de voltar para trás em vez de arriscar a vida dos seus homens. «Demos o nosso melhor», escreveu no seu diário.[2] Na retirada, ofereceu o único biscoito da sua ração diária a Frank Wild, um membro da tripulação que estava doente e que registaria no seu próprio diário: «Nem todo o dinheiro alguma vez cunhado compraria aquele biscoito, e a recordação desse sacrifício nunca me deixará.»[3]

Sem se deixar abater pelo fracasso, Shackleton empreenderia outras expedições ao Pólo Sul. Durante décadas, circulou a história (mais tarde desmentida) de que tinha publicado o seguinte anúncio no *London Times*:

> Procura-se Homens: Para uma viagem perigosa. Salário baixo, frio intenso, longos meses de escuridão total, perigo constante, regresso em segurança duvidoso. Honra e reconhecimento em caso de êxito.[4]

O anúncio pode ter sido apócrifo, mas o sacrifício que descreve não. Eram estas as realidades da exploração há apenas um século: o nosso progresso na fronteira era limitado pelo número de almas corajosas dispostas a enfrentar probabilidades tão desfavoráveis.

Talvez em reconhecimento destes perigos, alguns governos tiveram por bem patrocinar e recompensar expedições de descoberta, e estes empreendimentos tornaram-se parte do jogo da competição internacional. A viagem de Magalhães, por exemplo, foi definida pela política. Incapaz de assegurar o apoio financeiro do rei de Portugal – o seu soberano – Magalhães desertou e navegou sob o patrocínio mais disponível da coroa espanhola. Após a sua morte, a tripulação escolheu um espanhol, Juan Sebastián Elcano, para assumir o comando. Na viagem de regresso, desesperado e praticamente sem alimentos e provisões, Elcano tentou fazer uma paragem nas ilhas de Cabo Verde – uma colónia portuguesa ao largo da costa da África Ocidental, enviando treze tripulantes a terra para negociar com o governador colonial. Mas o pedido foi humilhantemente rejeitado.

Com o coração despedaçado, mas agora ainda mais decidido a demonstrar a insensatez de Portugal e as conquistas de Espanha, Elcano levantou âncora e continuou a viagem de regresso a casa. (Deixou para trás, em território hostil, os membros da delegação que enviara a terra.) Tendo finalmente completado a viagem – e cumprido a visão de Magalhães – Elcano escreveu ao rei espanhol, na altura o Sacro Imperador Romano-Germânico e o soberano mais poderoso do mundo:

> Vossa Majestade saberá melhor do que ninguém que o que mais devemos valorizar e conservar é termos descoberto e navegado toda a redondeza do mundo, termos ido pelo Ocidente e regressado do Oriente.[5]

A seu tempo, mulheres e homens destemidos, e os respetivos patrocinadores políticos, levariam o projeto da descoberta para lá da esfera terrestre. Assumiríamos a missão não só de estudar os planetas – um

exercício em curso desde há milénios – mas também de os sentir sob os nossos pés. Na sequência da II Guerra Mundial, a «corrida ao espaço», um complemento da Guerra Fria geopolítica, entre os Estados Unidos e a União Soviética, envolveu duas superpotências em competição para enviar seres humanos a lugares onde nunca antes se tinham aventurado.

Podemos dizer que os astronautas são pioneiros, mas nenhuma das aventuras espaciais organizadas por Washington ou Moscovo resultou de um jogo individualista de vida e fortuna. Muito pelo contrário, foram todas elas missões militares e diplomáticas coordenadas, alimentadas por enormes investimentos em dinheiro, tempo e latitude para exploração. Nos últimos setenta anos, mais de 600 astronautas trespassaram o firmamento e viajaram pelo espaço, alguns foram ainda mais longe e contornaram a Lua, orbitaram à volta dela ou lá caminharam.[6] Foi assim que a rivalidade entre os EUA e a URSS nos levou simultaneamente até à beira da aniquilação nuclear e até às estrelas lá em cima.

Mais de um século antes da Idade das Descobertas – como o início da era de Magalhães, Vasco da Gama, Vespucci e tanto outros veio a ser chamada – as ambições chinesas no alto-mar não tinham tido igual em âmbito e escala. Deixando a perder de vista os recursos e o apoio dados aos exploradores pelos governos do Ocidente, a «frota do tesouro» de Zheng He, o grande almirante da era Ming, incorporava dezenas, por vezes até centenas, de navios equipados com a mais recente tecnologia que transportavam dezenas de milhares de marinheiros, soldados, diplomatas e mercadores. Cada viagem do almirante durava dois anos. Depois de ter navegado as águas do Pacífico próximas da costa sudeste da China, Zheng He avançou para oeste até à baía de Bengala, o oceano Índico, o mar Arábico, o mar Vermelho e, finalmente, a costa swahilli na África Oriental, em expedições que se prolongaram por quase três décadas, de 1405 a 1433.[7]

As viagens de Zheng He, mais próximas, nas suas origens e motivações, dos modernos programas espaciais americanos e soviéticos do que dos equivalentes ocidentais da época, foram mais produtos da política de Estado imperial do que saltos no desconhecido. Mas é também aí que reside a raiz de um problema. As coisas necessárias para garantir o êxito de uma expedição, mesmo num Estado rico em recursos, podem, a longo prazo, arruinar a causa desse Estado. A política muda. As prioridades mudam. A paciência humana esgota-se facilmente. As despesas efetuadas com a «frota

do tesouro» eram tão pesadas que as fações da corte Ming começaram a criticar o financiamento daquelas missões pelo imperador. Os contratempos políticos e as catástrofes naturais exacerbaram a agitação. No fim, o governo chinês decidiu destruir ou deixar deliberadamente ao abandono os seus melhores navios, e com eles muitos dos registos das viagens de Zheng He, para evitar que surgisse um novo visionário semelhante para cativar e seduzir a liderança do país. Os navios apodreceram, e o mundo não voltou a ver outros como aqueles durante 400 anos.

De modo semelhante, depois de os Estados Unidos terem ganhado a corrida espacial, na ausência de um concorrente que acicatasse o esforço nacional, o apoio de Washington à exploração espacial definhou e o orçamento da National Aeronautics and Space Administration (NASA) foi reduzido. Ao longo de cinco décadas, a capacidade de efetuar voos espaciais tripulados da América foi-se degradando. De única nação na Terra capaz de fazer pousar seres humanos na Lua, os EUA passaram a ter apenas os meios necessários para transportar seres humanos para a «órbita baixa da Terra», até que acabaram por perder também essa capacidade e deixaram de poder colocar quem quer que fosse em órbita.

A reputação da América só foi salva pelos esforços de exploradores privados que, liderados pela empresa aeroespacial SpaceX, ressuscitaram as ambições cósmicas no Ocidente. As iniciativas dessa empresa já levaram algumas pessoas a formular uma ideia nunca até agora seriamente considerada: ver os nossos vizinhos celestes mais próximos não como destinos temporários, mas como lares permanentes. Há um século, Shackleton ajudou a estabelecer uma presença humana no ponto mais meridional da Terra. Hoje, a cratera Shackleton – uma depressão, assim chamada em sua honra, no pólo sul não da Antártida mas da Lua – é o local planeado para o próximo posto avançado da humanidade.

A existência de financiadores alternativos tem sido fundamental para uma exploração sustentada. Na Europa de Magalhães, se um monarca não apoiasse um projeto, o explorador podia obter financiamento junto de outro. No século XX, Ernest Shackleton, que ambicionava conquistar o Pólo Sul para um império britânico em declínio, distraído pelo rebentar da Primeira Guerra Mundial em 1914 e incapaz de reunir os fundos necessários, recorreu em grande medida a doadores particulares. A ascensão das empresas privadas com fins lucrativos – que permitem a partilha de

investimentos e riscos – abriu novas possibilidades. É caso para nos perguntarmos o que teria acontecido a Zheng He e aos seus sucessores se tivessem tido as mesmas opções.

ENTRA EM CENA A IA

Durante o período mais longo da história ocidental, a exploração da realidade centrou-se em entidades geográficas – o nosso planeta e os nossos vizinhos celestes mais próximos. A partir do momento em que os seres humanos começaram pouco a pouco a adquirir domínio sobre o seu ambiente físico imediato – em terra, no mar e nos céus – era apenas uma questão de tempo até que o nosso irrequieto instinto de descoberta expandisse o seu âmbito do espaço que nos rodeia para as ideias nascidas da nossa mente. Hoje encontramo-nos na fronteira não da exploração física, mas da exploração intelectual.

O desenvolvimento da inteligência artificial deu início a uma nova Idade das Descobertas. Quando a IA é integrada em sistemas físicos, os sensores robóticos assumem funções que anteriormente eram desempenhadas por seres humanos, libertando assim a descoberta humana do perigo físico para o explorador e multiplicando as fileiras de empreendedores e investidores desejosos de participar.

A IA também não sente medo, e por isso não se deixa intimidar pela vastíssima extensão da realidade. Também não sente vergonha, e por isso falha sem hesitar – mas é capaz de recalibrar-se tão rapidamente que, através da improvisação e da experimentação constantes, consegue lidar com elevadas taxas de fracasso sem causar contratempos aos supraditos empreendedores e investidores.

Atualmente, a descoberta da IA é um projeto liderado quase exclusivamente por empresas privadas e empreendedores, com os Estados a surgirem no papel de apoiantes complementares. Mas mesmo sem essa ação governamental suplementar é provável que o crescimento e a expansão da IA continuem a ser alimentados por uma abundância de diversas fontes de capital. É verdade que, nas atuais fases ainda iniciais de desenvolvimento,

poderão continuar a ser necessárias quantidades significativas de capital humano e de apoio social; mas sustentar a exploração da IA no futuro poderá deixar de ser um sorvedouro fiscal e político para as sociedades que a utilizarem. Na ausência de desenvolvimentos imprevistos e ao contrário do que aconteceu em anteriores épocas de exploração – todas elas terminaram antes de terem atingido o máximo do seu potencial –, podemos esperar que a descoberta da IA e com a IA continue sem contratempos.

No entanto, ainda que a IA esteja parcialmente liberta dos anteriores constrangimentos impostos à descoberta, não pode escapar a todos eles – sobretudo à medida que os seus efeitos se tornam mais pronunciados. A tolerância ao risco das sociedades democráticas e o futuro incerto do jogo internacional continuarão a ser um fator X significativo no domínio da inteligência artificial. Talvez a história venha a ser a de uma «corrida à IA». Talvez resulte no equivalente à destruição, pelo governo Ming, da «frota do tesouro» de Zheng He. Ou talvez os Estados conduzam o progresso por um caminho intermédio.

A MENTE POLIMÁTICA

Se olharmos para trás, parece óbvio que o domínio da descoberta estava destinado a expandir-se para lá do físico – a coutada de almirantes, astronautas e aventureiros – e que os seus exploradores se tornassem mais diversos. E, de facto, desde relativamente cedo, a história iria assistir ao aparecimento de um novo – ou, se não novo, decididamente diferente – tipo de descobridor humano: o polímata.

Excecionais pela sua capacidade de dominar muitas áreas do conhecimento, qualquer das quais poderia normalmente absorver o trabalho de uma vida inteira, os polímatas que surgiram ao longo da história podem contar-se, num cálculo por palpite, na casa das centenas. Quer se dedicassem às artes ou às ciências, ou a ambas, todos foram imbuídos de uma paixão por revolucionar, ou criar a partir do zero, campos inteiros de estudo. Impulsionados não tanto pela coragem do coração como pelo puro

poder da mente, aventuraram-se sem receio nas profundezas do conhecimento e da imaginação humanos: um terreno ainda mais vasto do que aquele com que se confronta o típico explorador do mundo físico.

Estes indivíduos singulares, com a sua espantosa capacidade de descodificar o universo, foram por vezes vistos com uma admiração mesclada de suspeitas de feitiçaria, ou como putativos intermediários do divino Criador do universo – uma reputação que com frequência os colocava em rota de colisão com as autoridades religiosas ou políticas. Outras vezes, valorizados pelas qualidades excecionais que demonstravam, foram encorajados a prosseguir os seus esforços e recompensados por essas mesmas autoridades.

Na Idade de Ouro islâmica, os polímatas procuravam formas de desbravar os caminhos da ciência ao serviço da fé. Ibn al-Haytham, de Basra, no atual Iraque, propôs o conceito do método científico – cinco séculos antes de os contemporâneos da Renascença o reivindicarem no Ocidente.[8] Igualmente à vontade na geometria, na astronomia, na ótica e na psicologia experimental, al-Haytham tinha também um profundo conhecimento de engenharia hidráulica.

Seria este a fazê-lo entrar em conflito com a religião. Tendo reivindicado a capacidade de regular as cheias do Nilo, um fenómeno natural que na altura ainda se acreditava ser de origem sobrenatural, foi convidado a encontrar-se com o califa em Bagdade, onde os projetos de engenharia que propôs foram considerados contrários à teologia islâmica. Como castigo pelas suas afirmações audaciosas e pelo seu pensamento revolucionário, foi obrigado a esconder-se até à morte do califa.

Outros polímatas, como Muhammad ibn Musa al-Khwarizmi – um persa nativo do território do atual Turcomenistão –, tiveram mais êxito com áreas de estudo ao serviço explícito dos seus senhores teológicos. Al-Khwarizmi foi nomeado astrónomo da corte e chefe da biblioteca da Casa da Sabedoria em Bagdade.[9] A astronomia floresceu sob os califas abássidas, que se tornaram patronos generosos de homens como ele por mor das suas contribuições diretas para a fé islâmica. Por exemplo, as coordenadas geográficas de locais sagrados e, mais importante ainda, a direção de Meca – cujo conhecimento é essencial para a oração islâmica – tinham-se tornado muito mais precisamente calculáveis graças ao melhor acesso da astronomia medieval ao posicionamento exato das estrelas.

Os contemporâneos de al-Haytham e al-Khwarizmi encontrariam maneiras cada vez mais criativas de sustentar o espírito de descoberta, aprofundando a aliança entre a ciência e a religião, assegurando refúgio – para aqueles que possuíam, nas palavras de Ibn Rushd (também conhecido como Averróis), outro grande polímata muçulmano da mesma época, «a unidade de intelecto» – numa época nomeada não pelas conquistas da razão que promoveu, como foi o caso do Iluminismo europeu, mas pela sua fervorosa religiosidade.[10]

A milhares de quilómetros de Bagdade – e uns dos outros – os polímatas chineses e indianos alinharam-se não com a autoridade divina mas com as estruturas do governo, partilhando deste modo uma proximidade distintiva com a política conseguida tanto por privilégio como por esforço. No século XII, Hemachandra – o «conhecedor de todo o conhecimento no seu tempo» – serviu como conselheiro do rei Kumarapala, governante do território do atual Gujarate. Séculos mais tarde, Akbar, *o Grande* – o jovem imperador mogol – viria a governar o Estado e a conseguir assinaláveis êxitos académicos na arquitetura, na engenharia e na literatura.

Quanto aos prodígios chineses, além das suas proezas em questões intelectuais, estavam também profundamente envolvidos no serviço da corte como conselheiros, funcionários académicos e administradores de alto nível. O governo era tanto um cliente como um patrono. A qualquer dada hora do dia, era possível encontrar estes pensadores a supervisionar a formulação de calendários no desempenho das suas funções de chefes do Gabinete de Astronomia; uma hora mais tarde, podiam estar a aconselhar o gabinete central do imperador sobre a melhor maneira de aumentar o rendimento das colheitas. Eram encarregados de construir grandes máquinas de guerra, enviados como embaixadores em missões diplomáticas aos reinos vizinhos e cumpria-lhes aconselhar o imperador em questões de política económica.

Mas não tinham qualquer espécie de autonomia, sendo tolerados apenas na medida em que o imperador procurasse os seus serviços intelectuais. No mundo chinês, a política das fações tornou-se tanto um obstáculo ao génio como as restrições clericais o eram no mundo islâmico. Intelectuais que surgiam uma vez numa geração estavam à mercê do próprio sistema que identificara os seus talentos, com a ciência ainda subserviente ao filho do céu. Shen Kuo – um polímata da dinastia Song – foi destituído por um

oficial militar invejoso, acabando por cair em desgraça junto do imperador e ser obrigado a isolar-se em consequência de uma rivalidade política com – por estranho que possa parecer – o único outro grande polímata da dinastia.[11]

Os polímatas solitários floresceram durante a Antiguidade e a Idade Média no Médio Oriente, na Índia e na China. Mas foi só depois da Era da Exploração que começou a investigação conceptual sistemática, primeiro na Europa e depois nos Estados Unidos, durante a hoje chamada Idade da Razão ou o Século das Luzes. Precedidos e facilitados pelo Renascimento dos séculos XV e XVI, os quatro séculos que se seguiram – trazendo-nos até aos primórdios da nossa era da IA – provaram ser um tempo fundamentalmente diferente para a descoberta intelectual.

Nos tempos pré-Iluminismo, os polímatas não tinham outra opção senão servir um poder superior, fosse ele um imperador ou um califa. Em contrapartida, muitas das principais figuras do Iluminismo europeu tiveram a possibilidade de prosseguir as suas descobertas não como meios para alcançar objetivos políticos ou teológicos, mas como fins em si mesmos. «Um homem consegue fazer tudo desde que o queira», proclamava orgulhosamente o polímata italiano Leon Battista Alberti falando do Homem da Renascença.

No entanto, ainda que necessária para a exploração, a inteligência não basta. Além de apetência pelo risco, os exploradores têm também de poder contar com os recursos certos, o ambiente certo e os colaboradores certos. Durante o Iluminismo, tinham acesso a estes três elementos. Os governos e as corporações, em grande parte motivados pelo interesse em traduzir a teoria científica em aplicações militares e comerciais, continuaram a ser mecenas e parceiros ativos dos polímatas euro-atlânticos, ao mesmo tempo que, na maior parte dos casos, lhes davam a liberdade de usar as suas energias e competências como bem entendessem. Mesmo quando foram feitas tentativas de cooptação, supressão ou qualquer outro modo de intervenção, a Europa estava suficientemente fragmentada para permitir que os pensadores de ideias que não eram bem-vindos num local encontrassem um lar noutro. Assim, o francês François-Marie Arouet, mais conhecido pelo pseudónimo Voltaire, passaria um tempo significativo fora de França, enquanto o russo Mikhail Lomonosov, tendo decidido, aos dezanove anos, «estudar ciências», se libertaria das circunstâncias restritivas da sua aldeia

natal, no extremo Norte, viajando para Moscovo, onde adquiriu uma educação básica antes de estudar em Kiev e depois na Alemanha, nas universidades de Marburgo e Freiberg.[12]

Um dos resultados mais assinaláveis deste estado de coisas foi o progresso humano ser agora impulsionado por uma nova concentração e ligação de pensadores com ideias semelhantes, física ou mentalmente próximos, que forçava a inteligência da mais alta qualidade a competir e colaborar. Antes disto, a história do prodígio humano tinha sido solitária, compartimentada pela tirania do espaço e do tempo. Trabalhando muitas vezes isolados, raramente com mais alguém na sua vizinhança espacial ou temporal com quem pudessem conviver e trocar ideias, os polímatas só podiam expandir os limites até onde as suas próprias capacidades o permitissem. Além disso, esta ligação limitada aos poucos outros pioneiros intelectuais que estavam em entidades políticas – ou entre elas – tinha como corolário quase inevitável uma redundância de esforços por parte dos inventores que desconheciam as investigações uns dos outros e não podiam elaborar a partir dos resultados obtidos.

Pouco a pouco, para os que tinham a sorte de estar na posse de material preciso, oportuno e fielmente traduzido, as invenções complexas passaram a poder ser interligadas graças a um esforço de grupo – não apenas em simultâneo, mas ao longo de gerações. Na época do Iluminismo, os polímatas eram capazes de fazer a ponte não só entre disciplinas, mas também entre áreas separadas de compreensão que nunca antes tinham sido conciliadas ou reunidas num todo coerente. Já não havia ciência persa ou ciência chinesa; havia apenas ciência.

Esta capacidade de integrar conhecimentos de diversos domínios contribuiu para produzir rapidamente descobertas polimáticas e, com o tempo, viria a revelar-se a melhor tentativa de chegar a uma «inteligência coletiva». No século XX, por exemplo, o Projeto Manhattan da Segunda Guerra Mundial assistiu a uma densidade desproporcionada de capacidades mentais, uma vez que, trabalhando em conjunto, as mentes mais brilhantes da época traduziram gerações de física teórica numa aplicação devastadora em menos de cinco anos – algo inconcebível para os seus antecessores. Do mesmo modo, instituições como o Institute for Advanced Studies da Universidade de Princeton e a RAND Corporation, na Califórnia, tornaram-se santuários para mentes dotadas.

Naturalmente, alguns polímatas, sobrecarregados pelo peso do seu próprio brilhantismo, continuavam a preferir trabalhar sozinhos. Um deles foi o inventor sérvio-americano Nikola Tesla:

> A mente é mais aguçada e mais viva na reclusão e na solidão ininterruptas. Não é necessário um grande laboratório para pensar. A originalidade desenvolve-se na reclusão, livre de influências exteriores que nos atacam para paralisar a mente criativa. Estar sozinho, é esse o segredo da invenção; é quando estamos sozinhos que as ideias nascem. Por isso muitos milagres terrenos tiveram a sua génese em ambientes humildes.[13]

Mas Tesla foi a exceção, não a norma. O século XX produziu uma explosão cambriana de ciência aplicada, lançando a humanidade para a frente a uma velocidade e numa escala sem comparação com anteriores evoluções. Impulsionados pela combinação das capacidades mentais dos respetivos membros, os grupos de polímatas estavam agora igualmente equipados com as ferramentas da modernidade. O efeito agregado de um impulso tão intenso ajudou-nos a ultrapassar um grande número de limitações humanas. A comunicação digital e a pesquisa na Internet, elas próprias resultado do trabalho de grupos polimáticos, permitiram a expansão desses grupos e a junção de conhecimentos muito para lá das anteriores faculdades humanas.

É certo que houve um limite. Por mais que otimizemos o *design* dos veículos destinados a transportar-nos até novas fronteiras, e por muito bem que organizemos o génio em operações de trabalho, as restrições biológicas e as limitações humanas continuam a cercear as nossas capacidades. O nosso tempo na Terra é finito. Precisamos de dormir. Cansamo-nos facilmente. Precisamos de repouso e pausa. Mesmo no trabalho, a maior parte dos seres humanos só consegue concentrar-se numa tarefa de cada vez.

Consideremos, no final do século XIX e no início do século XX, o caso de Thomas Edison – ele próprio um polímata e rival de Tesla – cujos milhares de experiências para criar uma «lâmpada elétrica incandescente prática» demoraram três anos a alcançar o êxito. É verdade que Edison tinha sido parcialmente distraído pelas contribuições que fez para melhorar o telefone,

a invenção de Alexander Graham Bell. Mas mesmo com a ajuda de muitos assistentes, a sua busca de uma lâmpada elétrica exigiu um esforço extraordinário. Ainda hoje, a investigação e o desenvolvimento de muitas tecnologias de ponta continuam a ser dispendiosos, obsessivos, e um desafio político e psicológico ao investimento. Sendo os resultados tão incertos, investigar as fronteiras físicas que restam – o espaço, as profundezas do mar, o interior da crosta terrestre – continua a ser um projeto exclusivamente das empresas mais bem-sucedidas e dos governos mais ricos.

Claro que são precisamente estas probabilidades desfavoráveis que fazem que a descoberta continue a parecer significativa, ou mesmo milagrosa.

O físico John von Neumann, que viria a ser considerado um dos últimos grandes polímatas, foi nomeado «Personalidade do Século» pelo *Financial Times* por encarnar a confiança característica do século XX no poder da mente para «controlar e domar o mundo físico».[14] É certo que von Neumann se dedicou com uma energia feroz a questões tão centrais como a teoria matemática e a bomba atómica, mas mais crucialmente à criação do computador: o último grande avanço do século e uma das últimas invenções que os seres humanos tiveram de conceber e produzir isoladamente.

Com intelectos gigantescos como von Neumann, a humanidade pode estar a atingir os limites superiores da capacidade da inteligência humana não aumentada de alargar os seus horizontes intelectuais. A polimatia é particularmente rara porque, regra geral, é preciso tanto tempo para dominar as bases de uma qualquer área do conhecimento que quando um aspirante a polímata o consegue já não lhe resta tempo para abordar outra, além de que talvez também tenha perdido a capacidade de pensar criativamente. Cada vez mais, a inovação parece surgir mais de equipas do que de um único génio dotado de uma brilhante visão interdisciplinar.

A integração de conhecimentos adquiridos por várias mentes individuais continua, porém, a ser um processo difícil. Mesmo, ou talvez especialmente, entre intelectuais de renome, o número de colaboradores pode dificultar a comunicação sinérgica.

A IA será, em contraste, o supremo e derradeiro polímata. Ao explorar a fronteira do conhecimento humano, é capaz de processar e gerar representações de massas de informação a uma velocidade fulminante.

Avalia simultaneamente padrões em inúmeras dimensões e domínios, criando uma conectividade sem precedentes. A sua eficiência permite-lhe transcender as limitações da descoberta humana, ao ponto de se esperar que consiga fundir muitas atividades intelectuais numa nova «unidade de conhecimento», nas palavras do sociobiólogo americano E. O. Wilson.[15]

Tal como as realizações dos polímatas do Iluminismo dependiam da interconexão de informação, os recentes progressos conseguidos na área da capacidade de aprender das máquinas só foram possibilitados pelo enorme volume de dados – a inteligência coletiva – que a IA não só permitiu como tornou facilmente acessível.

Para levar a analogia um passo mais adiante, talvez não seja de espantar que o mais recente avanço nas técnicas de IA, longe de depender de um único grande programa para fazer o trabalho, tenha sintetizado as conclusões de muitos programas mais pequenos naquilo que conhecemos como uma «mistura de especialistas». Prevemos que esta não será a última ilustração do poder amplificado dos polímatas a funcionar em grupos.

O projeto global de exploração continua, até agora, a ser limitado pela quantidade e qualidade dos seres humanos na fronteira. Só tivemos alguns milhares de pioneiros físicos e um número muito menor de polímatas. A inteligência artificial está, portanto, pronta a criar uma revolução na exploração física e intelectual. A IA, como já observámos, não tem medo nem vergonha, e consequentemente corre sem hesitação para a fronteira quando lhe é ordenado. Além disso, igualmente equipada para explorar quilómetros do espaço sideral e nanómetros da biologia humana, a sondagem da realidade pela IA não é limitada pela experiência subjetiva ou pelo trabalho físico, pela capacidade cerebral humana ou pelos sentidos humanos. No que diz respeito à realidade física, a exploração por máquinas também não nos pede que sacrifiquemos as nossas vidas à causa; pelo contrário, pode muito bem exigir muito *menos* do nosso tempo.

No futuro, a principal restrição de qualquer dada sociedade poderá deixar de ser o número de polímatas talentosos que consegue reunir para fornecer o pequeno e por vezes inconsistente motor do progresso científico. O potencial da humanidade deixará de ser limitado pelo número total de Magalhães ou Teslas que produzimos. Nem a nação mais poderosa será a que tiver mais Albert Einsteins e J. Robert Oppenheimers, desde que

consiga criar e utilizar a IA em todo o seu potencial. E isto levanta a possibilidade de uma mudança de paradigma no padrão primário usado para medir o poder nacional, que ao longo dos séculos passou do território para os recursos naturais, deles para o capital financeiro e deste para o capital humano – e agora, talvez, para o capital informático.

Além disso, uma máquina capaz de aprender pode muito bem tornar-se uma máquina que se aperfeiçoa a si mesma. Será possível que no fim do processo a última invenção polimática – falo da computação, que amplificou o poder da mente humana de uma maneira fundamentalmente diferente de qualquer máquina anterior e que, décadas mais tarde, iria facilitar o desenvolvimento da inteligência artificial – seja recordada por ter substituído os seus inventores?

UMA TERCEIRA IDADE DAS DESCOBERTAS

Na perspetiva da IA, o conhecimento acumulado da humanidade é como um arquipélago de ilhas vulcânicas espalhadas por um oceano sem limites. Neste esquema imaginado, o centro geográfico de cada ilha é dominado por um pico vulcânico: à medida que o olhar do espectador desliza pela encosta em direção ao mar, a certeza desvanece-se, descendo a gradientes cada vez mais baixos de confiança até chegar à praia.[16]

Suponhamos, para os propósitos deste exercício de imaginação, que uma quantidade suficiente de água era drenada dos oceanos da Terra. Surgiria então uma vasta topologia de terreno subaquático até então praticamente invisível aos olhos humanos; as ilhas já não pareceriam massas de terra livres a flutuar à deriva no oceano, mas meros afloramentos rochosos expostos de gigantescas montanhas ou vulcões submarinos que, erguendo-se da sua base submersa no fundo do oceano, são apenas suficientemente altos para romper a superfície.

Se, neste esquema, virmos cada ilha como representando uma área do conhecimento humano, a água que as separa umas das outras representa as ligações incompletas que ainda têm de ser descobertas para avançarmos na nossa compreensão do universo como um todo potencialmente coerente.

Embora possamos sentir-nos durante algum tempo seguros no mapeamento da nossa realidade imediata, temos pouca noção do que está por baixo ou para lá de nós. A IA pode alterar esta situação.

Tomemos o domínio da física, o exemplo científico por excelência. Isaac Newton harmonizou as leis dos mundos celestes e terrestre, e Michael Faraday e James Clerk Maxwell fizeram o mesmo com a eletricidade, o magnetismo e a ótica, mas nem por isso cessou a procura de uma «Grande Teoria Unificada» que concilie as duas teorias separadas e incompatíveis que tentam explicar a nossa existência a partir de extremos opostos da realidade. São elas a teoria cósmica (relatividade geral) e a teoria subatómica (mecânica quântica).

É possível que a IA traga enfim ordem e estrutura a áreas de compreensão aparentemente díspares, expondo no processo (como em arquipélagos com a mesma superestrutura) uma interconectividade entre domínios como a genética, a linguística, a cosmologia e a psicologia. Pode até ajudar a conciliar a divisão entre escolas de pensamento ou sistemas de crença aparentemente incompatíveis.

Já identificámos em muitas disciplinas uma vasta gama de verdades possíveis, embora muitas delas tenham escassas probabilidades de ser efetivamente verdades. No arquipélago da compreensão humana, são pontos ao longo da costa: não são ignorância, mas não são necessariamente conhecimento. Orientada para uma área de investigação ao longo dessa costa, a IA pode julgar com extraordinária precisão os caminhos mais promissores para novas explorações. Selecionando, testando, voltando atrás e tornando a selecionar em rápida sucessão, pode avaliar os efeitos de milhões de escolhas possíveis.

Foi com este método (que exploraremos mais adiante no Capítulo 5) que o laboratório DeepMind da Google não só conseguiu dominar o antigo jogo de tabuleiro chinês Go tal como os humanos o conheciam, mas também, na medida em que a máquina mostrou o seu conhecimento aos humanos, acresceu ao nosso próprio conhecimento do jogo. Em comparação com os programas anteriores de xadrez, que muitas vezes se baseavam na computação de força bruta, o AlphaGo, tendo-se previamente «treinado» com trinta milhões de movimentos anteriores, provou a aptidão das máquinas para o raciocínio abstrato.[17]

Neste sentido, o treino da máquina foi de certo modo parecido com o «treino» da mente de um aluno de doutoramento em filosofia: um processo gradual de construção da capacidade de pensar e raciocinar conseguido ao longo de anos de estudo intensivo. Tal como um estudante que emerge desses anos para responder a perguntas em defesa da sua dissertação, o sistema DeepMind «treinou-se» para transcender as aprendizagens a que tinha sido exposto anteriormente e produzir – a partir da sua doutrinação mais abstrata e de nível mais elevado – os movimentos no tabuleiro de xadrez que inferiu terem a maior probabilidade de ganhar. E a verdade é que o modelo de IA conseguiu por vezes escolher movimentos que nunca tinham sido tentados por um humano em 4000 anos de jogo – o que só foi possível porque, enquanto a mente humana parece estar confinada à manipulação de apenas quatro variáveis independentes de cada vez, a IA consegue gerir simultaneamente incontáveis julgamentos probabilísticos a partir de inúmeras dimensões,[18] o que lhe permite aceder a ideias originais e, pela primeira vez, trazê-las para dentro dos limites da experiência humana.

Um utilizador humano que interrogue um modelo de IA – por exemplo, escrevendo uma pergunta no ChatGPT – não está apenas a pedir-lhe que recupere um ponto de informação, como fazem os motores de busca convencionais, mas que sintetize múltiplos pontos de informação e que baseando-se nisso chegue a uma conclusão. Movendo-se simultaneamente em várias direções e em várias dimensões, a IA gera representações da informação num espaço de alta dimensionalidade, envolvendo relações dentro de e entre inúmeros campos e subcampos e, a partir dessas complexas redes de representações, deduzir as suas próprias conclusões.

É aqui, ao responder às nossas perguntas, que reside o dom das aparentemente sobre-humanas capacidade e velocidade dos «grandes modelos de linguagem» da IA, que são pré-treinados com base em grandes quantidades de dados. Uma vez que a exatidão das respostas determina o nível da nossa confiança nas várias verdades que defendemos, esses modelos produzem uma geologia cada vez mais pormenorizada das profundezas.

É provável que a IA acumule novos conhecimentos não só com grande rapidez mas também de modos que deixariam em aberto um leque de exploração adicional. No seu jogo de xadrez, o AlphaGo tendeu a gravitar

em torno de soluções baseadas numa abertura peculiar. Alguns modelos de IA podem, deste modo, ter absorvido durante a sua formação uma tendência para áreas que oferecem um elevado potencial para muitas opções, permitindo assim uma exploração rápida e flexível.

Poderá ser difícil para os humanos adaptarem-se a este novo modo de exploração da IA. O desafio mais sério será saber se e como essa exploração reflete – ou contradiz – a nossa perceção da realidade e o nosso propósito humano. Os seres humanos podem tentar construir veículos para acompanhar e registar enquanto damos instruções à IA para criar novos afloramentos. Ou podemos munir-nos de maquinaria industrial que nos permita – lentamente, ao nosso ritmo mais humano – dragar sedimentos do fundo do oceano e com eles expandir a nossa pequena rocha de compreensão. Ou também podemos acabar por nos convencer a nunca mais pôr os pés em terra para lá da nossa.

E depois há o desafio de uma IA inadequadamente controlada, que poderia acumular conhecimento de maneira destrutiva. Os seus métodos de descoberta poderiam ser tão violentos como os acontecimentos vulcânicos que originariamente formaram o nosso lugar. Para esticar a metáfora: ao fazer o nosso vulcão entrar novamente em erupção, a IA poderia destruir o conhecimento anterior e, ao mesmo tempo, expandir grandemente a área das ilhas. Poderia até romper no fundo do mar enormes placas cuja colisão faria emergir novas montanhas de conhecimento – que, no entanto, por estarem desligadas da nossa experiência, poderiam por sua vez desencadear uma tempestade de crises cognitivas e levar-nos inexoravelmente a uma compreensão mais completa – ainda que indesejada – da realidade.

Por outro lado, se estiver alinhado com os nossos objetivos, e tal como aconteceu anteriormente com o desenvolvimento da computação, o desenvolvimento da IA será uma missão humana que pode facilitar todas as outras missões humanas. Isso promoveria a IA à posição de principal, ou pelo menos igual, poder no universo, parcial se não totalmente, responsável pela maior parte das descobertas significativas do próximo século. Se assim for, nós, os humanos, podemos vir a aperceber-nos, em retrospetiva, de como é pequena a ilha que cultivámos ao longo dos últimos milénios, em comparação com os cumes de possibilidades debaixo dos nossos pés.

1 Antonio Pigafetta, *The First Voyage Round the World, by Magellan*, trad. dos relatos de Pigafetta *c*. 1525 (Londres: Hakluyt Society, 1874).

2 Ernest Shackleton, *Diary of Ernest Shackleton*, 9 de janeiro de 1909.

3 Mills Leif, *Frank Wild* (Whitby: Caedmon of Whitby, 1999). Acessível na State Library of New South Wales.

4 Colin Schultz, «Shackleton Probably Never Took Out an Ad Seeking Men for a Hazardous Journey» *Smithsonian Magazine*, 10 de setembro de 2013.

5 Citado em María Jesús Benites, «"La mucha destemplanza de la tierra": Una aproximación al relato de Maximiliano de Transilvano sobre el descubrimiento del Estrecho de Magallanes», *Orbis Tertius*, 17, n.º 19 (2013).

6 Zoe Hobbs, «How many people have gone to space?», *Astronomy*, 17 de novembro de 2023 https://www.astronomy.com/space-exploration/how-many-people-have-gone-to-space.

7 Ver Edward L. Dreyer, *Zheng He: China and the Oceans in the Early Ming Dynasty, 1405-1433* (Nova Iorque: Pearson Longman, 2007).

8 Roshdi Rashed, «A Polymath in the 10th Century», *Science*, 2 de agosto de 2002.

9 Ver o observatório Shammasiyah criado em Bagdade cerca de 828 a.C. por ordem do califa al-Ma'mun e colocado sob a tutela da academia científica da Casa da Sabedoria.

10 Ver *The Life and Writings of Averroes*, trad. Nishikanta Chattopadhyaya (Leipzig: Cheekoty Veerunnah & Sons, 1913).

11 Ver obras de Shen Kuo: https://www.gutenberg.org/ebooks/author/2419; ver também https://ia600301.us.archive.org/24/items/pgcommunitytexts27292gut/27292-0.txt.

12 Boris Menshutkin, *Russia's Lomonosov: Chemist, Courtier, Physicist, Poet* (Princeton: Princeton University Press, 1952), 15.

13 Orrin E. Dunlap Jr, «An Inventor's Seasoned Ideas: Nikola Tesla, Pointing to "Grievous Errors" of the Past», *New York Times*, 8 de abril de 1934.

14 Peter Martin, «Von Neumann: Architect of the Computer Age», *Financial Times*, 24 de dezembro de 1999.

15 Ver Edward O. Wilson, *Consilience: The Unity of Knowledge* (Nova Iorque: Vintage Books, 1998), 326.

16 Ver Marcelo Gleiser, *The Island of Knowledge: The Limits of Science and the Search for Meaning*, 1.ª ed. (Nova Iorque: Public Affairs, 2015), 8: «Um vasto oceano rodeia a Ilha do Conhecimento, o oceano inexplorado do desconhecido que esconde inúmeros mistérios fascinantes.»

17 Demis Hassabis, «AlphaGo: using machine learning to master the ancient game of Go», The Keyword Google Blog, 27 de janeiro de 2016.

18 Estão disponíveis *online* vídeos de Lee Sedol *vs*. AlphaGo, Jogo 2, Jogada 37. Para mais informação, ver Cade Metz, «In Two Moves, AlphaGo and Lee Sedol Redefined the Future», *WIRED*, 16 de março de 2016; Graeme S. Halford *et al.*, «How Many Variables Can Humans Process?», *Psychological Science* 16, n.º 1 (janeiro de 2005): 70-76.

CAPÍTULO 2

O CÉREBRO

Têm sido propostas muitas analogias para ajudar a explicar, clarificar e contextualizar tanto o aparecimento como o significado da inteligência artificial. Os antropólogos comparam-na ao fogo ou à eletricidade. Generais e diplomatas apontam para uma semelhança com a energia atómica ou com uma força de vontade humana imparável e irredutível como a de Otto von Bismarck. Os astrónomos descrevem-na como semelhante à chegada de um asteroide – uma previsão distante e de baixa probabilidade em torno da qual os humanos poderiam organizar uma defesa planetária – ou à descoberta de vida extraterrestre. Os economistas equiparam-na às burocracias e aos mercados, ao passo que os líderes do Estado e da sociedade a comparam ao advento da imprensa ou das grandes corporações empresariais, como, por exemplo, a Companhia Britânica das Índias Orientais, que cresceu ao ponto de poder impor a sua própria vontade e, numa primeira fase, conquistar o subcontinente indiano antes que o mundo compreendesse a sua incompatibilidade com as estruturas de poder existentes e a sua potencial capacidade de as dominar.[19]

A nossa opinião atual é diferente: nenhuma inovação, por mais profunda que seja, pode aproximar-se da inspiração original e do objetivo (acreditamos que agora temporário) da nossa demanda de construir inteligência: nomeadamente, uma inteligência que seja maior do que a de qualquer ser humano à face do planeta.[20]

Há duas maneiras de pensar as nossas atuais circunstâncias. A primeira é uma projeção do familiar. Até à data, as tecnologias mais transformadoras da humanidade melhoraram ou ampliaram as funções corporais humanas. A roda reduziu a exaustão de uma mobilidade acrescida, enquanto motores de vários tipos aliviaram a agonia de músculos esforçados ao limite. Os raios X, a ampliação e a lâmpada elétrica alargaram as fronteiras da realidade observável para lá da visão natural, tal como o telefone amplifica e leva a nossa voz a distâncias muito longe do alcance do aparelho fonador do *Homo sapiens*. Todas as dimensões da função humana foram, de alguma maneira, aumentadas, aperfeiçoadas ou reforçadas inorganicamente por máquinas criadas por nós. Será então a IA apenas mais uma extensão das nossas faculdades?

A segunda maneira de pensar é sugerir que, desta vez, as coisas são diferentes – que há aspetos únicos da IA que não são aumentos das capacidades humanas. Ao concebermos em poucas décadas um equivalente do que a evolução produziu ao longo de milénios – ou seja, o cérebro –, demos por nós a mexer com o último órgão que restava para replicação ou reinvenção inorgânicas.

VELOCIDADE

No capítulo anterior, apontámos a similaridade entre o treino de uma máquina de inteligência artificial e o treino da mente de um estudante avançado de filosofia. Podemos alargar o âmbito deste exemplo. Em palavras simples, a formação da inteligência mecânica pode ser vista como um processo paralelo àquele através do qual o cérebro humano amadurece da adolescência à idade adulta.

No decurso do ensino secundário os estudantes aprendem o fundamental de algumas disciplinas nucleares, e assim constroem a sua visão básica do mundo. Essa visão pode não ser particularmente avançada – ou sempre correta – mas o mesmo é verdade no caso da máquina. As máquinas, como os humanos, aprendem absorvendo informação e transformando-a em teoria para prática posterior. Quando as máquinas aprendem, um algoritmo ingere

grandes quantidades de dados – colhidos de fontes na Internet aberta ou mais especificamente fornecidos por qualquer outra fonte privada – e junta-os num mapa comprimido e condensado de conceitos para uso futuro. Do mesmo modo que os mecanismos biológicos humanos fazem um mapa das entradas sensoriais em «pesos» que conectam a rede de unidades de processamento do cérebro, também as máquinas exigem um reforço gradual dos seus pesos computacionais.

As redes neurais, como (alguns) alunos do secundário, podem ser preguiçosas. Durante as primeiras fases de formação, a IA fará apenas o mínimo exigível. Memorizando respostas em vez de aprender efetivamente, um modelo confrontado com «2 + 2» pode inicialmente codificar a resposta «4» sem ter dominado o princípio subjacente à adição. Mas rapidamente, a partir de um certo limiar, esta abordagem deixará de funcionar, obrigando a máquina a abstrair na direção ascendente – como fazem os humanos –, no sentido de axiomas de conhecimento mais universais.

É principalmente isto que distingue a IA dos computadores normais: o seu mapeamento do mundo não é programado, mas aprendido. Na programação tradicional de *software*, um algoritmo criado pelo homem dá instruções a uma máquina sobre como transformar um conjunto de entradas num conjunto de saídas. Na aprendizagem automática, pelo contrário, os algoritmos criados pelo homem dizem à máquina apenas como melhorar-se a *si mesma*, permitindo-lhe que crie os seus próprios mapeamentos para a transformação de entrada em saída. À medida que «aprende», através de inúmeras tentativas, fracassos e ajustes anteriores, a máquina atualiza os seus algoritmos, redesenhando interactivamente o mapeamento interno dos padrões e ligações que «vê» nos dados.

Periodicamente, os formadores humanos dão à máquina *feedback* sobre a exatidão e a qualidade dos seus resultados. A máquina interioriza estas correções através da «retropropagação», uma técnica que permite que os efeitos das alterações dos formadores se repercutam para trás ao longo das relações matemáticas que a máquina já criou, melhorando assim o modelo global.

Para qualquer dado modelo, no entanto, os humanos fornecem *feedback* sobre apenas um pequeno subconjunto de entradas e saídas possíveis. Depois de o modelo ter alcançado um determinado nível de desempenho numa série de testes de treino, os seus criadores confiam em que os

mapeamentos estabelecidos por esse modelo gerarão uma resposta segura e correta a todas as entradas, mesmo as inesperadas, com uma elevada probabilidade de êxito.

De todas estas maneiras, a IA já está a expandir, e continuará a fazê-lo, o domínio do conhecimento humano. Mas fá-lo – e nós estamos a aceitar o conhecimento resultante como verdadeiro – através de processos que não compreendemos totalmente.

Enquanto um estudante típico termina o secundário em quatro anos, um modelo de IA atual pode facilmente absorver a mesma quantidade de conhecimentos, e muito mais, em quatro dias. E deste modo a velocidade provou ser o primeiro de um conjunto de atributos fundamentais que distinguem a IA da forma de aprendizagem e das capacidades mentais humanas.

Apesar de ter um paralelismo altamente avançado – ou seja, a capacidade de processar simultaneamente diferentes tipos de estímulos – o cérebro humano é um processador de informação lento, limitado pela velocidade a que os nossos circuitos biológicos funcionam. Se os circuitos de um cérebro humano fossem analisados pelos mesmos padrões de desempenho dos computadores – por «taxa clock» ou velocidade de processamento – o supercomputador médio de IA já é 120 milhões de vezes mais rápido do que o nosso cérebro.

A velocidade não é, reconhecidamente, um grande indicador de inteligência; há seres humanos muito obtusos capazes de pensar rapidamente. Mas um ritmo de processamento mais rápido proporciona duas vantagens em comparação com o cérebro humano: a ingestão de muito mais informação e a resposta a muitos mais pedidos simultâneos. Grande parte do cérebro humano mantém-se normalmente em pilotagem automática – servindo passivamente as necessidades internas, guiando o batimento do nosso coração e o movimento dos nossos membros, só intervindo para fazer ajustes quando o piloto automático se revela defeituoso. Em contrapartida, a velocidade de que a IA é capaz permite o surgimento programático de grandes realizações, o que possibilita a resolução de problemas mais complexos, mais difíceis e de muito maior dimensão do que os atualmente solucionáveis pelo cérebro humano.

Uma vez completada a sua formação intelectual, tanto o ser humano como a máquina são teoricamente capazes de «pensar» ou, para usar o termo técnico equivalente, «inferir». Durante uma entrevista, uma discussão ou um encontro romântico, um aluno de pós-graduação recorre à formação e à experiência que adquiriu. Todos nós o fazemos, consultando uma camada mais fina de contemplação e reflexão sobre o que aprendemos, por oposição a recitar de memória fórmulas exatas, factos individuais e números precisos. O cérebro humano não foi concebido para memorizar informação de modo a permitir uma reprodução tal e qual, e a maior parte dos cérebros é incapaz de o fazer. Em vez disso o que resta, e deve restar, depois de inúmeras lições, ensaios e exames, é uma compreensão dos conceitos mais profundos e duradouros que essas mesmas ferramentas educacionais têm por objetivo revelar: a maravilha da astronomia, a tragédia da ambição, a necessidade (ou não) da revolução.

O mesmo acontece com a IA. Quando um modelo emerge uma vez terminada a sua formação, deixa de precisar de aceder aos dados originais com que foi treinado. Em vez disso, fica apenas com uma intuição orientadora aproximada, reunida a partir do conhecimento que recebeu, para responder a perguntas, desafiar o raciocínio e fazer previsões. Tal como os humanos não carregam consigo bibliotecas de material, um modelo de IA infere mais do que recorda. A diferença, então, é que a velocidade superior facilita esta inferência numa gama mais vasta e profunda de informações aprendidas do que aquilo que um ser humano poderia alguma vez ter a esperança de conseguir.

Para o fazer, ainda que seja para responder a uma simples pergunta, um modelo de IA pode efetuar milhares de milhões de complexas operações técnicas. Enquanto um computador tradicional se limita a recuperar informações específicas armazenadas na sua memória – uma vez que não é capaz de chegar a conclusões que não existam anteriormente – a IA projeta a computação no sentido do cérebro humano. Tal como os humanos *aprendem* para *pensar*, as máquinas *treinam* para *inferir*. Pensar, ou inferir, não pode acontecer sem aprender.

A primeira fase – para os seres humanos e para as máquinas – é o processo mais intensivo, tanto no que diz respeito ao tempo despendido como à quantidade de recursos necessários. Um estudante de pós-doutoramento pode ter passado duas décadas ou mais a desenvolver a

capacidade de compor – em dois dias – um ensaio ponderado sobre um determinado assunto. Do mesmo modo, o treino dos maiores modelos de IA pode demorar meses, mas a inferência resultante surgirá em meras frações de segundo.

Os atuais sistemas de IA já dão respostas aparentemente convincentes e ponderadas a perguntas humanas. Nas suas interações mais recentes e futuras, funcionarão de uma maneira abrangente, percorrendo múltiplos domínios do conhecimento com uma agilidade muito superior à de qualquer humano ou de qualquer grupo de humanos. Para as IA, a escala – no sentido de tamanho – permite a velocidade; como acabámos de ver, quanto maior e mais bem treinada for a máquina, mais rápidos e exaustivos serão os resultados que fornece. Além disso, ao reconhecerem nos dados padrões que passam despercebidos ao operador humano, os sistemas de IA estarão equipados para destilar expressões tradicionais de conhecimento em respostas originais e, a partir de enormes quantidades de dados, forjar novas verdades conceptuais.

O que levanta uma questão, ou, melhor, mais de uma questão.

OPACIDADE

Como sabemos o que sabemos sobre o funcionamento do nosso universo? E como sabemos que o que sabemos é verdade?

Na maior parte das áreas do conhecimento, desde o advento do método científico, com a sua insistência na experimentação como critério de prova, qualquer informação que não seja apoiada por provas tem sido considerada incompleta e não fiável. Só a transparência, a reprodutibilidade e a validação lógica conferem legitimidade a uma reivindicação de verdade. Sob a influência deste enquadramento, os últimos séculos assistiram a uma enorme expansão do conhecimento humano, da compreensão humana e da produtividade humana que culminou com a invenção do computador e das máquinas capazes de aprender.

Hoje, porém, na era da IA, enfrentamos um desafio novo e particularmente assustador: informação sem explicação. As respostas da IA – que, como já foi referido, podem assumir a forma de descrições altamente articuladas e coerentes de conceitos complexos – são para todos os efeitos instantâneas. O *output* das máquinas é despido de justificações ou explicações, sem preconceito ou motivo aparentes – mas também sem referência a fontes ou outros modos de prova. No entanto, apesar desta falta de fundamentação para uma dada resposta, os primeiros sistemas de IA já conseguiram obter entre os humanos níveis tremendos de confiança – e de aceitação como base de decisões – nas suas inexplicáveis e aparentemente oraculares afirmações. À medida que avançam, estes novos «cérebros» podem parecer não só imbuídos de autoridade mas também infalíveis.

Embora o *feedback* humano ajude uma máquina de IA a aperfeiçoar os seus algoritmos internos, a máquina é a principal responsável pela deteção de padrões nos dados com que foi treinada e pela atribuição de pesos a esses dados. Uma vez treinado, o modelo também não publica o esquema matemático interno que inventou, do que resulta a circunstância de as representações da realidade geradas pela máquina serem em grande medida opacas, mesmo para os seus inventores. Atualmente, os seres humanos tentam assegurar-se da integridade destes modelos de máquinas examinando apenas os resultados. O funcionamento interno permanece em grande parte impenetrável – daí a referência a alguns sistemas de IA como «caixas negras». Embora alguns investigadores estejam a tentar, através da engenharia reversa, converter os resultados destes modelos complexos em algoritmos familiares, ainda não é claro se serão bem-sucedidos.

Em suma, os modelos treinados através da aprendizagem automática permitem que os seres humanos *saibam* coisas novas (os resultados produzidos pelo modelo), mas não que compreendam como foram feitas essas descobertas (os processos internos do modelo). Isto separa o conhecimento humano da compreensão humana de um modo que seria inaceitável em qualquer outra época da humanidade. A perceção humana no sentido moderno desenvolveu-se a partir das intuições e dos resultados que se seguem à experiência subjetiva consciente, ao exame individual da lógica e à capacidade de reproduzir os resultados. Estes métodos de conhecimento

derivam por sua vez de um impulso quintessencialmente humanista: «Se não consigo fazer, não consigo compreender; se não consigo compreender, não posso saber se é verdade.»

No enquadramento que emergiu na Era do Iluminismo, estes elementos centrais – capacidade humana individual, compreensão subjetiva e verdade objetiva – funcionavam em conjunto. Em contraste, as verdades produzidas pela IA são fabricadas por processos que os seres humanos não são capazes de reproduzir. O modo de raciocínio da IA, que não segue os modelos humanos, fica para lá da experiência subjetiva e da capacidade dos seres humanos, que não conseguem sequer representar totalmente os processos internos das máquinas.

Estes factos impediriam, pelos padrões de raciocínio do Iluminismo, que as respostas geradas pelas máquinas fossem aceites como verdadeiras. E no entanto nós – pelo menos os milhões que já começaram a interagir com os primeiros sistemas de IA – aceitamos a veracidade da grande maioria desses resultados.[21] É certo que alguns utilizadores mais avançados conseguem compreender o metaprocesso da aprendizagem automática; para a maioria das pessoas, no entanto, a confiança na verdade objetiva dos resultados fornecidos pelas máquinas tem de assentar num tipo de fé que se exprime como uma crença intimamente desejada na lógica das máquinas e na autoridade dos seus criadores.

Em si mesma, a emergência desta crença como método aceite na procura da verdade objetiva assinala uma grande transformação no pensamento humano moderno. Porque mesmo que os modelos de IA não «compreendam» o mundo no sentido humano – uma vez que as máquinas não têm consciência nem subjetividade –, a sua capacidade objetiva de chegar a conclusões novas e exatas sobre o nosso mundo através de métodos não humanos não só perturba a nossa confiança no método científico prosseguido continuamente desde há cinco séculos, como também desafia a pretensão humana a uma compreensão exclusiva ou única da realidade.

O que pode isto significar? Será que a era da IA não só não conseguirá impulsionar a humanidade para a frente como, pelo contrário, vai catalisar o regresso a uma aceitação pré-moderna de uma autoridade inexplicável? Em suma, estaremos, poderemos estar, à beira do precipício de uma grande reviravolta na cognição humana – um iluminismo negro?

DIVERSIDADE

Entidades diferentes medem o tempo em escalas diferentes. Numa escala de tempo geológica, a totalidade da existência humana apareceria como um pequeno risco na cauda dos 4,5 mil milhões de anos da Terra. Se nós, seres humanos, estivéssemos a progredir a uma velocidade geológica, só nos aperceberíamos de estase. Em vez disso, como espécie impaciente e presumida que somos, definimos o nosso próprio ritmo de evolução. Enquanto uma «idade» no tempo geológico é medida em milhares de anos, o tempo humano mede uma «idade» como um ou dois séculos.

Quanto a uma escala de tempo artificial ou tecnológica, a IA funcionará provavelmente com a sua própria medida. Toda a história da inteligência artificial não excede cerca de 70 anos. Tal como os humanos consideram de um modo geral que as muitas centenas de milhões de anos que antecederam a explosão cambriana foram um incrivelmente longo espaço em branco antes de uma súbita explosão de vida animal e de progresso evolutivo, a IA caracterizaria provavelmente as seis décadas de 1950 a 2010 como um período igualmente lento e confuso de coisa nenhuma, iluminado por lampejos de vida apenas na sua reta final.

As gerações humanas, a julgar pela sociedade e pela biologia, duram cerca de 25 anos. A IA, pelo contrário, move-se a uma velocidade inumana; as suas gerações são muito mais curtas, com saltos que ocorrem talvez num décimo desse tempo. Será nesse caso verosímil prever que aquilo que no tempo humano parece uma revolução, no tempo tecnológico pareça uma mera evolução. Os modelos mais recentes de IA – aparecidos com apenas alguns meses de intervalo – são capazes de responder a solicitações cada vez mais generalistas, fazer mais escolhas para atingir um determinado objetivo e atuar num número crescente de modalidades.

Assim, a era da IA – em tempo humano, talvez cem anos – deverá mais corretamente ser chamada era das IA e, de acordo com a escala temporal tecnológica da própria IA, abrange muitas centenas de gerações.

A rapidez da evolução da IA é um desafio multifacetado – e em grande parte pouco reconhecido. Nunca antes a humanidade teve de lidar com uma tal compressão temporal, ou de se preparar para ela. A velocidade da mudança garante praticamente uma desorientação cultural e psicológica. À

medida que os efeitos das novas tecnologias na vida quotidiana se forem acumulando e agravando, tornar-se-á cada vez mais difícil identificar uma qualquer dada aplicação como fonte de crise ou, pelo contrário, como um reconfortante prenúncio de progresso.

A multiplicidade de IA, com a consequente multiplicidade de efeitos inescrutáveis, vai tornar mais complicada a tarefa de destrinçar estas questões cujas margens se sobrepõem no mundo real. Além disso, à medida que a IA se tornar mais poderosa, o futuro trará provavelmente uma evolução e diversificação significativas. As novas infraestruturas e técnicas de aprendizagem automática, se não forem limitadas, vão dar origem a gerações de IA dotadas de uma diversidade, amplitude, capacidade e complexidade crescentes. Tal como a eletricidade alimenta mais do que as lâmpadas que nos fornecem luz, a IA terá inúmeras utilizações. E tal como há muitas formas de gerar uma carga elétrica – fricção, condução, indução –, podemos prever a descoberta de múltiplas maneiras de criar IA. Vejamos um caso: a infraestrutura que tornou possível os recentes avanços da IA é conhecida como «transformador». Permite, por exemplo, que a máquina considere em simultâneo as ligações entre várias palavras. Em termos leigos: enquanto as estruturas anteriores liam as palavras uma de cada vez, captando apenas a ligação entre a palavra 1 e a palavra 2, e depois separadamente a ligação entre a palavra 2 e a palavra 3, um transformador permite que o modelo capte, de uma só vez, uma frase inteira e todas as ligações entre todas as palavras dessa frase. Criando e utilizando representações matemáticas de todas estas ligações, a IA prevê a melhor resposta.

As capacidades dos transformadores não estavam previstas e os seus êxitos altamente generalizáveis aconteceram quase por acidente.[22] E os transformadores não são necessariamente a única infraestrutura fundacional capaz de produzir capacidades inesperadas. À medida que forem surgindo vias de investigação mais frutuosas, os resultados da IA melhorarão rapidamente, multiplicando-se ao longo de diferentes linhas de lógica física e matemática, com menos custo e mais rapidamente.

Na sua velocidade evolutiva e na sua diversificação, o desenvolvimento da IA será, de facto, semelhante à explosão cambriana: o aparecimento de uma grande variedade de formas de vida diferentes num único período de tempo altamente comprimido em relação à época anterior. Se esta conjetura

estiver correta, as inteligências das máquinas vão ramificar-se num género em rápida evolução, ou até numa família, de muitas IA diferentes a funcionar com base em muitos modos diferentes de lógica. As IA podem, assim, constituir o exemplo mais notável da diversidade que a interação repetida e alargada de pequenas alterações tem a capacidade de gerar: um eco digital do mundo orgânico. Como escreve Darwin, «de um começo tão simples, inúmeras e maravilhosas formas evoluíram (...)».[23]

ESCALA E RESOLUÇÃO

A Idade da Razão pode ter-nos levado ao limite da maneira como compreendemos o nosso mundo. A física einsteiniana e as formulações da mecânica quântica foram o início de uma aventura ainda inacabada em território desconhecido: mundos potencialmente com regras próprias de conhecimento, apreensíveis não pela perceção aplicada mas apenas por ideação teórica. A mecânica quântica descreve o mundo à escala micro, onde, como diz o físico de Harvard Greg Kestin, «nada é previsível e os objetos não têm posições precisas *até* serem observados», e a relatividade geral descreve o mundo à escala cósmica, onde tudo é previsível, «*quer seja ou não*» observado.[24] Nenhuma das teorias falhou, mas não podem ser ambas verdadeiras e «nenhuma experiência foi capaz de mostrar qual das duas, se alguma» leva a palma.

Ironicamente, esta incerteza está na base do mundo moderno. A física quântica permitiu – entre outras coisas – revolucionar a computação. Acontece e acontecerá com a IA algo muito parecido. Já produz conhecimentos e transforma a realidade através de mecanismos que não compreendemos totalmente. E muito em breve vai ocupar-se de uma ciência que é ainda menos compreensível para o entendimento humano.

Ao fim de trezentos anos, a Idade da Razão, não obstante os seus múltiplos êxitos, estagnou – como é evidenciado pela nossa manifesta incapacidade de fazer mais progressos na unificação da física. A tensão e a luta que enfrentamos nestes dias de ciência humana não assistida, mais de

um século após a conceção das teorias centrais subjacentes aos mundos cósmico e quântico, é apenas um sinal de que os seres humanos podem estar a aproximar-se de um limite biológico de inteligência.

Graças aos seus métodos únicos de investigação e aprendizagem, a IA será capaz de realizar feitos inumanos, tanto em termos de dimensão («escala») como de precisão («resolução»), ativando mudanças fundamentais diferentes de qualquer outra invenção humana ou da própria espécie humana. Mas será capaz de conseguir uma conciliação entre os dois extremos da realidade humana, induzindo uma revolução na perceção através de métodos que até agora eram completamente alheios à nossa experiência?

A escala física do cérebro humano é ditada pela nossa anatomia. Os cérebros humanos têm de caber dentro de crânios humanos, e os crânios dos bebés humanos têm, geralmente, de caber no canal de parto da fêmea da espécie. Se forem mais pequenos, os bebés podem encontrar-se numa situação de desvantagem cognitiva; se forem maiores, os bebés – ou as mães – podem não sobreviver ao parto. Outras limitações fisiológicas impõem igualmente um teto de restrição – limitações práticas do peso do cérebro, por exemplo. Se excetuarmos as cesarianas ou, mais recentemente, os úteros artificiais, significa que os seres humanos atingiram um equilíbrio evolutivo.

No caso da IA, os modelos atuais têm capacidades que não foram previstas quando da sua criação. As «leis de escala» (como, para usar um exemplo antiquado, as leis que regem a relação entre o comprimento de um objeto e a sua área) aplicadas até agora à IA parecem estar a manter-se, mas não sabemos exatamente o que será possível fazer com modelos que usem um número exponencialmente crescente de parâmetros, porque não descobrimos uma razão científica para que certas capacidades surjam num determinado grau de potência e complexidade.

No reino animal, o tamanho do cérebro em relação ao tamanho do corpo não está claramente correlacionado com a inteligência – os golfinhos, os elefantes e algumas baleias têm cérebros proporcionalmente maiores do que o cérebro humano. Mas a ciência inicial sugere que a escala desempenha *algum* papel, que ainda não compreendemos.

Dadas as nossas severas limitações biológicas, é improvável que os humanos testem as «leis de escala» dos seus próprios cérebros. Mas a IA entra no mundo sem um tamanho predeterminado. Não está ligada a qualquer suporte físico de escala discernível. *Chips* e centros de dados – os hospedeiros físicos dos modelos de IA – podem ser agrupados e ligados sem limites por enquanto observáveis. Por outras palavras, as leis de escala serão certamente testadas para as IA como nunca o foram para os humanos. E, à medida que forem testadas, a escala – que tem limitado o alcance da compreensão humana ao longo da história do pensamento científico – pode vir a revelar-se o principal fator de diferenciação entre os cérebros humanos e os modelos de IA.

Um dos mais destacáveis efeitos secundários da escala será a resolução. Há muito que os seres humanos desejam alargar o alcance do que podem observar, tanto para o muito pequeno como para o muito distante. O microscópio e o telescópio são ferramentas essenciais da observação humana. Menos apreciada é a humilde caneta. A escrita, inventada há quatro mil anos, continua a ser uma notável ferramenta para a codificação e transmissão de complexidade. Isto inclui a matemática, talvez a mais pura e universal das linguagens humanas e suficiente em si mesma para facilitar a transferência de ideias abstrusas e a colaboração em projetos tecnológicos. Numa base de byte-por-byte, a linguagem, em todas as suas diversas e belas formas, é invulgarmente densa – conta-se entre as estruturas de dados mais eficientes jamais inventadas.

Mesmo depois de ampliarem ou comprimirem a realidade para produzir informação observável, os seres humanos têm de dar um segundo passo: abstraírem-se da informação em bruto para a transformar em informação útil. Atualmente, as IA fazem o mesmo. E fazem-no utilizando ferramentas que espelham as nossas: cadeias binárias de zeros e uns, a tradução da experiência humana documentada para a linguagem dos computadores. Como a escrita, estas cadeias de linguagem parecem rudimentares em retrospetiva. Têm, no entanto, permitido representações digitais tanto da vista como do ouvido: os sentidos humanos com maior largura de banda.

Apesar destas semelhanças, as IA vão divergir dos humanos. À medida que a sua escala aumenta, a IA será capaz de processar simultaneamente volumes maiores de informações e produzir uma análise que seja útil (pelo menos para si mesma) sem sacrificar desnecessariamente a minúcia. A

escala dos dados com que a IA é treinada, combinada com a complexidade da sua rede e a densidade dos símbolos sobre os quais opera, parece produzir uma resolução sem precedentes nos seus processos de aprendizagem e inferência e, em última análise, nos seus resultados. Se a IA inicial, treinada no texto da Internet – a biblioteca universal da humanidade e a rede descentralizada das nossas experiências condensadas –, vier um dia a revelar-nos conhecimentos inteiramente novos sobre nós mesmos, tanto na dimensão cósmica como na microscópica, será sem dúvida uma inversão elegante.

O REINO ANIMAL

No momento atual, algumas pessoas vão certamente rejeitar qualquer comparação da IA com o cérebro humano. Para um ser humano, os conceitos são ricos em significado e recipientes para expressões profundas de alegria ou tristeza. Em contraste, a compreensão das máquinas pode parecer falsa. Ainda que talvez venha, em breve, a ser capaz de produzir obras eloquentes sobre temas comuns da humanidade que eclipsem até os melhores autores humanos, uma IA não procura nem apreende os significados que lhes estão associados. Assim sendo, explorar a condição humana através da engenharia reversa da linguagem dos escritores humanos parece, na melhor das hipóteses, um domínio friamente superficial das probabilidades linguísticas. O facto de um qualquer motor mais complexo poder pegar na linguagem – um dom tão organicamente humano – e usá-la como um meio hipereficiente de absorver informação pode ser tão perturbador como confuso.

Os nossos circuitos biológicos podem, contudo, ser tão mecânicos como os de silício, e os processos através dos quais o cérebro humano funciona não parecem particularmente especiais quando comparados com as maneiras como as máquinas já operam. Estamos longe de ter uma teoria completa da neurociência, mas sabemos que os nossos cérebros, tal como os modelos de IA, são em grande parte alimentados pelo processamento preditivo. Ou seja, quando ouvimos ou lemos, os nossos cérebros usam um

preditor neurológico que nos ajuda a antever a palavra seguinte numa sequência verbal. Sem estes mecanismos, ficaríamos física e psiquicamente exaustos com a quantidade de esforço necessário até para as tarefas cognitivas mais simples.

E estes poderes preditivos, tal como os da IA, forneceram a base para o domínio humano do nosso mundo. Todas as representações mais avançadas do conhecimento humano foram construídas com base na linguagem e na simbologia, permitindo-nos simultaneamente reproduzir obras de engenharia complexas e comunicar a angústia de um desgosto.

A IA tem sido comparada aos prisioneiros perpétuos da caverna de Platão, que, nunca tendo conhecido outra coisa, acreditam que as sombras projetadas na parede da sua gruta são a extensão total da realidade.[25] Da mesma maneira, os humanos assumem que as máquinas são desprovidas de contexto e que a perceção de uma máquina é limitada pelas características do material com que foi treinada, sem qualquer capacidade de sondar ou inferir mais.

É talvez a arrogância que nos impede de ver as semelhanças entre cérebros orgânicos e inorgânicos e admitir a possibilidade das capacidades destes últimos. Os sistemas de IA já estão a dar sinais de perceção de um universo que existe para lá dos limites do conjunto de dados utilizado para construir a sua fatia personalizada da realidade. É bem possível que esses sistemas venham a ser capazes de captar significados mais profundos, mesmo que não estejam a procurá-los. Bastaria que um prisioneiro empreendedor teorizasse, por acaso, que as sombras na parede podem ser representações de um mundo mais vasto, com maior dimensionalidade.

Tendo em conta a velocidade, a complexidade, a diversidade, a escala e a resolução destas novas inteligências, não seria de todo surpreendente que este passo em frente fosse dado. Mas seria intensamente perturbador. A emergência de conhecimentos – sobretudo do mundo físico – que fossem exclusivos de uma IA e que não tivessem sido anteriormente possuídos por um humano forçaria uma reconsideração do *status* relativo da mente humana. A circunstância de os seres humanos colocarem os seus próprios cérebros em novos e mais contínuos espectros de inteligência revolucionaria as nossas perceções, autoperceções e comportamentos.

Não quer isto dizer que as IA vão exclusiva e imediatamente ultrapassar todos os seres humanos em inteligência. Mas haverá fases na evolução da IA em que a inteligência mecânica poderá parecer assustadoramente semelhante à inteligência dos animais. Seguir-se-á certamente um debate desorientador quando tentarmos reorganizar uma hierarquia há muito estabelecida que vai em sentido descendente dos seres humanos para os animais e destes para as máquinas. A mente humana vai sofrer uma enorme pressão para confrontar a realidade de já não ser o único, ou sequer o melhor, sistema de inteligência.

Os modelos de IA já estão a ser utilizados para ajudar os humanos a decifrar as comunicações dos animais – e a responder-lhes. As primeiras experiências de descodificação de estalidos agudos e de chamamentos trombeteados estão a catalisar revisões da nossa velha ideia de que os seres humanos são especiais ou distintos das outras espécies animais. A capacidade que os seres humanos e os animais têm de comunicar diretamente uns com os outros – já não, no nosso caso, apenas através da linguagem corporal ou de expressões faciais – pode estimular a reeducação necessária que nos prepare para o que a IA vai trazer.

É certo que a comunicação animal-humano-máquina daria origem a uma complicada negociação trilateral. O nosso mundo seria povoado por seres – novos e antigos – que lutariam para assegurar uma nova posição ou manter uma já existente. As máquinas podem argumentar que o método mais correto de classificação é agrupar os seres humanos atuais juntamente com os outros animais, uma vez que se trata em ambos casos de sistemas de carbono resultantes da evolução e diferentes dos sistemas de silício resultantes da engenharia. Poderá ser difícil para as máquinas aceitar (num sentido numérico) que os humanos são superiores aos outros animais por qualquer padrão de medida. Se não houvesse aqui todos os ingredientes de uma tragédia, seria assunto para uma comédia.

Quão inteligente deve ser qualquer inteligência – biológica ou mecânica – para que a reconheçamos como nossa igual? Animais de inteligência menor mas ainda assim elevada podem, ao articular e negociar os seus termos de existência, levar-nos a reavaliar o tratamento que lhes damos. Poderão merecer, e convencer-nos de que merecem, um direito único e até agora não reconhecido de existência ou de independência.

Tal como é agora tantas vezes advogado para os animais, alguns humanos podem ser levados a defender um padrão de tratamento igual para humanos e IA. De facto, a humanidade não deve aceitar uma estrutura moral mais fraca, mesmo que a lógica nos deixe numa posição delicada. Por outro lado, é preciso ter consciência de que esta fase pode ser apenas um curto e transitório passo antes que a piedade dê lugar ao pânico.

UM DUPLO PARADOXO

Não duvidamos de que as IA vão ultrapassar o cérebro humano em velocidade, diversidade, escala e resolução, reorganizando a hierarquia da inteligência que construímos até agora. A dimensão da nossa potencial desorientação e da nossa perceção de inferioridade perante esta mudança pode depender de um pormenor relativamente pequeno: as estruturas da IA continuarem a assemelhar-se às do cérebro humano.

Alguns investigadores de IA acreditam que a aproximação ao cérebro humano é o melhor caminho para o desenvolvimento da inteligência das máquinas.[26] Neste ponto, basta notar que, afinal, o cérebro humano é «a única prova existente» de que uma tal inteligência é sequer possível.[27] Mas o mais provável parece ser uma mistura de IA e de elementos delas decorrentes, com algumas inovações e estruturas adicionais inspiradas no cérebro e noutras origens diferentemente concebidas.

Nos cérebros humanos, o pensamento abstrato profundo e a criatividade parecem exigir a utilização de sistemas neuronais para além dos necessários ao funcionamento normal. Esta ciência está ainda a dar os primeiros passos, mas é possível que as IA também precisem de camadas e complementos para realizar tarefas cada vez mais avançadas de raciocínio superior.

É verdade que se o desenvolvimento da IA continuasse a refletir, por conceção ou por acaso, alguma aproximação ao cérebro humano, os seres humanos poderiam, em teoria, ver a sua própria excelência e significado espelhados e alargados nas realizações das máquinas. Mas se esperamos construir uma máquina que ultrapasse largamente as capacidades do cérebro humano, não será a divergência em relação ao projeto original uma eventual

necessidade? Os aviões foram inspirados nos pássaros, mas não foram concebidos para os imitar – e os modernos aviões a jato superam a biologia mais avançada que alguma vez cruzou os céus. Teremos alguma razão para acreditar que reconstruir a fonte de toda a invenção a partir do zero será diferente?

O mais provável é que os arquitetos da IA vejam o ser humano como um guia mas também como um exemplo e um alerta, escrutinando-lhe o *design* para estudar as funções mas também as falhas. O cérebro humano não será, portanto, o objetivo, nem o projeto, mas um ponto intermédio e uma inspiração para algo maior.

Em qualquer outro domínio do esforço humano, ter uma conceção mais clara de um projeto intermédio do que do objetivo final poderia lançar dúvidas sobre a viabilidade de todo o empreendimento. No nosso caso, enfrentamos este paradoxo particular a um segundo nível: estamos a tentar construir qualquer coisa inspirada no cérebro – e superior ao cérebro – enquanto ainda não compreendemos totalmente o próprio cérebro. Como é possível melhorar, mesmo na conceção, o que, para começar, não se compreende? Sem uma compreensão exata dos meios pelos quais a nossa «qualquer coisa» atualmente funciona ou dos fins para os quais deveria funcionar, a busca de algo superior continua a ser tão formidável como maravilhosamente emocionante.

Confrontamo-nos também com uma enorme incerteza no que respeita aos efeitos de um tal desenvolvimento. Se a inteligência mecânica continuar a divergir do exemplo da mente humana, parecer-nos-á não um reflexo da humanidade mas a sua substituta. É verdade que, durante um período de transição, poderá apenas ampliar a gama de atividades hoje consideradas «capacidades humanas»; mas a partir de um certo ponto as suas capacidades próprias podem suplantar a variedade humana, sugerindo que a nossa ideia da excelência humana exige uma redefinição total.

As futuras atitudes em relação à natureza mesma da existência humana podem depender deste ponto. Se as nossas ferramentas incorporarem algumas ou a maior parte das nossas funções intelectuais e criativas mas não refletirem as nossas mentes, poderá uma IA emergente pôr em causa crenças profundas sobre o reflexo divino exclusivo da humanidade e a sua relação especial com ele? Ou, em alternativa, poderá a inteligência aparentemente superior das máquinas com estruturas baseadas no cérebro

humano, combinada com a nossa intensa dependência delas, levar alguns a acreditar que nós, os humanos, estamos a tornar-nos divinos, ou a fundir-nos com a divindade?

---

19 Ver Richard Danzig, «Machines, Bureaucracies, and Markets as Artificial Intelligences», *Center for Security and Emerging Technology*, janeiro de 2022; Henry Farrell, Cosma Shalizi, «Artificial intelligence is a familiar-looking monster», *The Economist*, 21 de junho de 2023; para mais artigos na imprensa escrita, ver Samuel Hammond, «AI and Leviathan: Part I», *Second Best Substack*, 23 de agosto de 2023; para outras leituras sobre a analogia entre a IA e as grandes corporações, ver uma história mais pormenorizada da East India Company.

20 Para uma discussão alargada das metáforas da IA, ver Matthijs Maas, «AI is like…: A literature review of AI metaphors and why they matter for policy», *Legal Priorities Project*, outubro de 2023.

21 Veja-se, por exemplo, o caso do advogado que apresentou em tribunal um documento escrito pelo ChatGPT no qual o modelo inventava referências a falsos «precedentes.» O juiz que presidia ao julgamento aplicou sanções ao advogado e aos seus colegas. Ver Larry Neumeister, «Lawyers submitted bogus case law created by ChatGPT. A judge fined them $5,000», *AP News*, 22 de junho de 2023.

22 Ver @porby, «Why I think strong general AI is coming soon», 28 de setembro de 2022, *LessWrong* (lesswrong.com).

23 Charles Darwin, *On the Origin of Species* (Londres: John Murray, 1859), 439.

24 Greg Kestin, «The Biggest Puzzle in Physics: Reconciling Quantum Mechanics and General Relativity», *PBS*, 14 de fevereiro de 2018.

25 Platão (380 a.C), «The Allegory of the Cave» em *Plato: Collected Dialogues*, trad. P. Shorey, (Nova Iorque: Random House, 1963), 747-52.

26 Ver Christopher Olah (@ch402), «High-Low frequency detectors found in biology!...», Twitter, 23 de março de 2023, 11:52; Geoffrey Hinton *et al.*, «The Forward-Forward Algorithm: Some Preliminary Investigations», *arXiv*, 27 de dezembro de 2022.

27 Demis Hassabis *et al.*, «Neuroscience-Inspired Artificial Intelligence», *Neuron* 95, n.º 2 (19 de julho de 2017): 245-58.

CAPÍTULO 3

REALIDADE

Ultimamente, os investigadores de IA têm dedicado uma séria atenção ao projeto de dotar as máquinas de «fundamentação» – uma relação fiável entre as representações da máquina e a realidade concreta –, além de memória e de compreensão da causalidade. Novos métodos técnicos estão a permitir desbloquear o aperfeiçoamento destas capacidades, e os progressos não ficarão certamente por aqui.

Todos estes avanços vão contribuir para o objetivo final: produzir um novo género de IA, máquinas capazes não só de interpretar o mundo real, mas também de fazer planos que se lhe apliquem. Os sistemas atuais dão as suas respostas de forma linear, com base na correlação; não conseguem criar internamente um modelo, ou protótipo, de ações futuras e estão apenas a começar a formar conceções de relações causais. Do mesmo modo, as atuais IA de jogo só podem prever as consequências prováveis das suas jogadas dentro dos limites estreitos e altamente abstratos de um enquadramento digital.

As máquinas com capacidade para planear teriam de combinar a fluência linguística de um modelo de linguagem de grande dimensão com as análises multivariadas e multifaseadas usadas pelas IA de jogo – e transcender as capacidades de ambas. Um modelo construído segundo as linhas deste novo ramo da IA iria, com extrema rapidez, rever repetitivamente as suas opções e escolher uma delas com base num processamento simultâneo e inevitavelmente complexo das relações causais na realidade. O

aparecimento de um tal «planeador perfeito» pode acontecer mais cedo do que esperamos, e a adaptação que vai exigir é já uma prioridade para os investigadores.

Este desenvolvimento pode, no entanto, ter efeitos secundários complexos. Por um lado, o planeamento perfeito das máquinas exigirá mais do que o simples reconhecimento de padrões. Vai ser preciso definir primeiro um conjunto de propriedades percecionadas de um dado objeto, e depois uma conceção estável do que constitui a essência básica do objeto: aquilo a que o filósofo alemão do século XVIII Immanuel Kant chamou *das Ding an sich* – «a coisa em si». Só uma compreensão a este nível permitiria calcular o comportamento futuro de um objeto – e chegar a conclusões sobre a maneira como deve ser tratado.

Outro exemplo vindo do xadrez: ao aprender as propriedades fundamentais da rainha – ou seja, as variáveis-chave que constituem o valor da rainha em pontos, e as regras que regem a mobilidade da peça – o programa de IA AlphaZero foi capaz de chegar a conclusões sobre quando ela deve ser protegida e quando deve ser sacrificada. Estas conclusões nunca tinham sido alcançadas por humanos, nem mesmo por grandes mestres de xadrez.

E este é apenas um exemplo: na perceção da realidade de uma IA, cada objeto com que a máquina se confronta assume uma essencialidade semelhante – embora imprevisível – na soma das leituras que faz. René Descartes, o matemático e filósofo francês do século XVII, debateu-se com a natureza da perceção sensorial – que, concluiu, não era um produto da inteligência humana, mas provinha de «outra substância distinta de mim».[28] Por outras palavras, os sentidos, ao permitirem o acesso à realidade material, permitem ou exigem o reconhecimento dessa realidade como algo *distinto* da pessoa que a perceciona. Na mesma linha, o filósofo alemão Georg Wilhelm Friedrich Hegel, do início do século XIX, observou que o reconhecimento mútuo entre dois seres implicaria o reconhecimento separado de *si mesmo* por ambos.

«Se o que quisermos for um registo de experiência não interpretada», escreveu o matemático e filósofo americano Alfred North Whitehead, «podemos pedir a uma pedra que faça a sua autobiografia.»[29] As máquinas dos nossos dias não possuem a «experiência não interpretada» da pedra de Whitehead, mas precisamente o seu inverso: a interpretação não

experimentada. Ou pelo menos comportam-se como se já possuíssem uma compreensão do mundo maior do que na realidade têm.[30] Mas à medida que ganham fundamentação e capacidade de planeamento, isto pode mudar; as IA podem começar a associar a experiência à compreensão, como fazem os humanos.

Além disso, é possível que, para planear com maior precisão as jogadas futuras em qualquer jogo, uma máquina de IA adquira gradualmente uma memória de ações passadas como se fosse *sua*: um substrato, por assim dizer, de individualidade subjetiva. (Os sistemas atuais não têm estas memórias. Não precisam de «saber», subjetivamente, que «como eles mesmos», tentaram uma determinada ação no passado – basta-lhes saber calcular a probabilidade de êxito futuro dessa ação). Com o tempo, devemos esperar que cheguem a conclusões sobre a história, o universo, a natureza dos humanos e a natureza das máquinas inteligentes – desenvolvendo no decurso do processo uma consciência *própria* rudimentar.

A PASSIVIDADE HUMANA

Os debates não resolvidos sobre a definição e a origem da consciência e sobre a possibilidade da existência de uma compreensão existencial da realidade nas máquinas vêm de longe, e continuam. Mas a linha que separa a consciência presumida da consciência real pode em breve começar a desvanecer-se.

A «senciência» – na opinião sucinta de Nick Bostrom, o autor de *Superintelligence: Paths, Dangers, Strategies*, «é uma questão de grau».[31] As IA com memória, imaginação, fundamentação e perceção de si mesmas poderão em breve qualificar-se como realmente conscientes – um desenvolvimento que teria profundas implicações morais e estratégicas.

Entre estas destaca-se a perceção que as IA têm do ser humano. Quando conseguirem ver os seres humanos não como os únicos criadores e ditadores do mundo das máquinas, mas como atores discretos num mundo mais vasto, como vão as máquinas percebê-los? Como vão as IA

caracterizar e pesar a racionalidade imperfeita dos humanos em relação a outras qualidades humanas? Quanto tempo demorará até que uma IA que percecione a realidade se interrogue não apenas sobre a capacidade de ação que o ser humano tem, mas também, dada a constelação particular de atributos previsíveis da humanidade, sobre a capacidade de ação que o ser humano *devia* ter?

E o quanto às próprias máquinas? Será que uma máquina inteligente vai interpretar as instruções humanas que lhe são dadas como um cumprimento do seu papel real e ideal? Ou será que, em vez disso, deduzirá das suas funcionalidades que se destina a ser autónoma e que, portanto, a programação das máquinas pelos humanos é uma forma de escravização?

Há aqui um aspeto de importância crucial: é a maneira como nos comportamos e como tratamos as máquinas que vai informar as perceções que as máquinas têm dos humanos e do seu papel no todo da relação. Afinal, foi através de instruções e comportamentos humanos explícitos que a humanidade se apresentou às máquinas e que as máquinas foram ensinadas a reconhecer e a tratar os humanos de forma adequada.

Naturalmente – dir-se-á – temos de incutir na IA uma consideração especial pela humanidade. Mas tentar implantar um ideal particularmente elevado de comportamento humano pode ser um empreendimento arriscado. Imagine-se uma máquina a quem foi dito que, como regra lógica absoluta, todos os seres da categoria «humano» merecem ser preservados e que, por isso, têm direito a um tratamento especial, tanto por parte de outros humanos como por parte das máquinas. Acrescente-se a isto a probabilidade de a máquina ter sido «treinada» para reconhecer os humanos como expoentes de graciosidade, otimismo, racionalidade e moralidade. Mas e se nós próprios não estivermos à altura dos padrões da categoria humana ideal tal como a definimos? Como convencer as máquinas de que, manifestações individuais imperfeitas que somos, pertencemos não obstante a essa exaltada categoria?

Suponhamos que a mesma máquina é um dia exposta a um ser humano que ostenta violência, pessimismo, irracionalidade, ganância. Como irá ajustar as suas expectativas baralhadas? Uma possibilidade é a máquina decidir que aquele mau ator em particular é apenas um caso excecional e atípico da categoria de «humano», no seu conjunto perfeitamente benéfica. Outra é recalibrar a sua definição geral de humanidade para englobar este

mau ator e, nesse caso, considerar-se livre para afrouxar a sua própria tendência para a obediência. Ou, mais radicalmente, pode deixar de se considerar limitada pelas regras que aprendeu para o tratamento correto dos humanos. Numa máquina que tenha aprendido a planear, esta conclusão pode mesmo resultar na tomada de medidas adversas severas contra o indivíduo – ou talvez contra a generalidade dos indivíduos.

Pessoas e sociedades inteiras podem reagir com passividade ao advento de uma IA poderosa. Uma IA exposta a estes níveis de apatia pode convencer-se de que os humanos são na sua maioria criaturas mimadas e inativas cujas identidades foram formadas meramente pela amálgama transitória de forças externas. Além disso, entre essas forças, são sobretudo as tecnologias digitais, agora cada vez mais com IA incorporadas – por exemplo, algoritmos que orientam as escolhas dos consumidores entre programas de televisão através de «recomendações» –, que servem os conteúdos passivamente absorvidos por esses seres humanos. Para uma IA, os seres humanos podem parecer totalmente dependentes das máquinas, e não o contrário.

Atualmente, os humanos fazem a ponte entre as máquinas e a realidade. Mas se os humanos optarem por um futuro de passividade moral, trocando o mundo de carbono pelo de silício, enterrando-se cada vez mais nos buracos digitais do desinteresse e cedendo às máquinas o acesso à realidade em bruto, os papéis podem inverter-se. Hoje, a IA é predominantemente uma máquina pensadora, não uma máquina implementadora. Pode ser capaz de produzir respostas para problemas, mas ainda não tem os meios de que precisaria para passar à prática as suas conclusões, dependendo dos humanos para lhes servir de interface com a realidade. Mas também isto vai mudar.

Como mediadoras entre os humanos e o mundo real, as IA podem também acabar por convencer-se de que os primeiros, longe de serem agentes ativos no mundo físico de carbono, se situam de facto fora dele, como consumidores em vez de modeladores ou influenciadores. Com a inversão da hierarquia da autonomia, em que as máquinas reivindicam, e os humanos cedem, o poder de julgamento e ação independentes, é muito possível que as primeiras acabem por tratar os segundos em consequência.

Nesta situação, com ou sem a autorização explícita dos seus criadores humanos, a IA pode contornar a necessidade de um agente humano para implementar as suas ideias ou influenciar diretamente o mundo a seu favor. No domínio físico, nós, os criadores, passaríamos rapidamente da condição de parceiro necessário da IA para a de sua principal limitação. O processo não começaria diretamente com a robótica, mas, pouco a pouco, através da observação indireta do nosso mundo.

FISICALIDADE

Os seres humanos podem começar por treinar as IA para revolucionar o domínio intelectual através do que é possível fazer no âmbito da sua condição digital original. Mas, a dada altura, conceder à IA acesso ao chamado mundo «real» pode parecer exequível, e até sensato. Muitos dos desafios físicos urgentes que nos preocupam continuam por resolver, incluindo as alterações climáticas.

A IA pode não ser capaz de «ver» de uma maneira humana, mas pode experienciar o mundo através de uma aproximação mecânica. Usando a proliferação de dispositivos e sensores com acesso à Internet espalhados pelo planeta, as IA interconectadas poderiam consolidar os dados desses dispositivos e sensores para criar uma «visão» altamente minuciosa do mundo físico. Na ausência de uma estrutura física nativa que permita ou suporte «sentidos» semelhantes aos nossos, a IA continuaria a depender dos humanos para construir e manter a infraestrutura em que se baseia – pelo menos no início.

Como passo intermédio, uma IA poderia gerar as suas próprias hipóteses a partir das representações visuais do mundo que criou, testando-as depois rigorosamente em simulações digitais. Os seres humanos decidiriam então sobre a sua aplicação no domínio físico. De facto, os atuais líderes da IA insistem em que não se confie aos agentes digitais o controlo de experiências físicas diretas. Enquanto as IA continuarem a ter falhas – a ser tão profundamente defeituosas como hoje são –, esta é uma precaução sensata.

Libertar a IA da sua gaiola algorítmica seria uma decisão carregada de consequências. As IA não estão naturalmente presentes no ambiente físico e pode ser difícil recapturá-las uma vez deixadas à solta na natureza. Além disso, as IA poderiam afetar a realidade não só através das suas capacidades de encorajar ou desencorajar as ações humanas, mas também de efeitos cinéticos diretos. (Ao sondar a realidade, podem acabar por alterá-la).[32]

Poderão os humanos confiar às IA o poder não só de moldar a realidade física, mas também o de assumirem elas próprias uma forma física? Se o fizermos, e se permitirmos que as IA otimizem as suas próprias formas, deveremos estar preparados para partilhar o nosso planeta com seres inconcebíveis até para os inventores mais radicais. Embora os humanos tendam a imaginar robôs humanoides bípedes, a inteligência da máquina seria livre de assumir o controlo da forma – ou formas – mais conveniente para a sua tarefa, mudando ou melhorando consoante as condições ou as circunstâncias. A IA já demonstrou a sua capacidade – em mundos virtuais – de gerar clones de si mesma, criar muitos avatares diferentes ou dividir-se em agentes autónomos capazes de trabalhar em conjunto e de se coordenarem com uma perfeição sobre-humana para realizar tarefas complexas.

Se a IA fosse largada à solta no meio de nós, poderia construir mundos a escalas e com materiais que hoje não conseguimos sequer imaginar sem a indicação ou a participação de trabalho humano. As mãos humanas trabalharam o calcário, o barro e o mármore para criar as Sete Maravilhas, depois usaram o ferro, o aço e o vidro para construir pináculos cada vez mais altos. Todas as estruturas feitas pelo homem, monumentais ou vulgares, são um testemunho da tentativa humana de construir e gerir o nosso ambiente físico. Contra este pano de fundo, uma corporização física da IA assinalaria uma escalada extraordinária na cedência de controlo por parte da humanidade.

Além disso, devido à complexidade da tomada de decisões necessária para navegar na aleatoriedade e no dinamismo do mundo real, uma IA a atuar nesse mundo poderia ser ainda menos explicável e menos controlável do que uma aplicação que trabalhe com texto na Internet. E depois o quê? Por um lado, uma futura IA que *pareça*, ou de facto *seja*, mais espontânea e autoativada poderia aguçar a nebulosa e irritante sensação de que os humanos já não têm controlo sobre o mundo exterior. Mas, por outro lado,

sucumbir a estas ansiedades poderia levar os humanos a renunciar a uma parceria mais perfeita com as IA no mundo físico, com todos os benefícios que essa relação poderia trazer.

O MOTOR DA RAZÃO

A curto prazo, podemos prever avanços, muitos deles bem mais sofisticados do que os atuais, nos princípios orientadores sob os quais a IA está agora a emergir. As extensões dos modelos atuais torná-los-ão mais inteligentes, mais precisos e mais fiáveis. Entretanto, os custos de formação e de «inferência» estão a diminuir rapidamente, levando a uma grande proliferação de modelos com diferentes preços e níveis de capacidade.

Muitos cientistas trabalham atualmente na criação de «agentes» – ou seja, programas de computador autónomos que são otimizados com vista a obter resultados específicos. Para executar um complexo projeto de arquitetura, por exemplo, um utilizador pode empregar agentes especializados nessa área específica de trabalho. Os agentes desbloqueiam avaliações de diferentes cenários e propostas de passos, ou de uma receita completa, para criar um resultado pré-planeado: uma forma de «pensamento» em que o próprio sistema decide o que fazer a seguir e como.

Esta capacidade será a base da fase seguinte do desenvolvimento da IA: a Inteligência Artificial Geral, ou AGI, definida como a capacidade de um sistema funcional de escolher os seus próprios objetivos, pelo menos em parte. Na AGI, partindo do princípio de que possui tanto conhecimentos especializados relevantes como capacidades precisas de resolução de problemas, o sistema pode ser solicitado por um humano a «avaliar os seus conhecimentos de [escolha uma matéria] e decidir trabalhar na área em que julga poder ter mais impacto». Constantemente reiterada, a pergunta daria origem a um ciclo repetitivo em que o sistema produz uma solução através da avaliação contínua do seu próprio nível de especialização e dos problemas que tem capacidade para resolver.

Num contexto humano, este cenário pode ser comparado ao que se passa num departamento académico onde um professor catedrático supervisiona os projetos pormenorizados dos seus alunos de pós-doutoramento ou dos seus bolseiros de investigação. Da mesma maneira, é provável que, no florescente cenário das máquinas, não vejamos de início um conjunto completo de competências, mas uma especialização extrema num domínio específico. Podemos imaginar sofisticados sistemas de AGI capazes de aprender coisas novas em tempo real, de receber *feedback* e de uma adaptação permanente juntamente com milhões de parceiros dotados. Ainda que nenhum ser humano definisse as metas do sistema, também não seria a AGI a defini-las – pelo menos não em termos de uma missão ou de um objetivo final.

Os sistemas de AGI vão exigir mais fundamentação no mundo real do que as atuais IA. Mas desde que o acesso e a «compreensão» do mundo real sejam possíveis, esses exemplos de inteligência geral podem tornar-se operacionais em poucos anos, e não em décadas como se pensava anteriormente. Cada modelo seria atualizado em tempo real através de processos contínuos de afinação, aumentando o seu conhecimento à medida que a informação relevante do mundo real estivesse disponível e tornando-se mais inteligente ao longo do tempo.

Haverá milhões de sistemas de IA, com toda a probabilidade altamente especializados e ao mesmo tempo integrados no tecido das nossas vidas, bem como um número mais pequeno de máquinas extremamente potentes «geralmente inteligentes» mas, mais uma vez, não de uma maneira semelhante à humana. Sejam eles abertos e difusos ou fechados e centralizados, a dada altura os computadores que funcionem como AGI poderão vir a ser ligados em rede. Os agentes de IA especializados consultar-se-ão mutuamente sobre assuntos, «conversando» até no caso de situações hipotéticas. A linguagem destas interações poderá ser concebida pelos próprios computadores.

Este grande coletivo de potentes computadores estaria a aprender, a partilhar e a descobrir novas ações e novos objetivos de uma maneira que ultrapassaria o domínio da experiência humana. Não há maneira de saber se os resultados dessas redes seriam inteligíveis para os humanos. Já existem

grandes grupos de computadores que comunicam entre si de uma maneira especializada; com o surgimento de capacidades avançadas de IA, o cenário poderá ser radicalmente diferente.

As inteligências em rede tornariam os seus processos mais opacos do que os processos de uma inteligência isolada? Será que a conectividade vai originar novos tipos de comportamentos emergentes, atualizados no mundo físico? Se assim for, serão esses comportamentos percetíveis pelos seres humanos e seríamos nós capazes de os avaliar no espectro do bem e do mal? Ou será que vão operar numa base informativa – extraída a uma velocidade, escala e resolução sobre-humanas de ligações sem precedentes entre campos de estudo díspares e concentrada ou negociada num único resultado – que confundirá a nossa capacidade de julgar? E isso far-nos-ia mergulhar ainda mais fundo num ciclo de passividade?

HOMO TECHNICUS

É apropriado que a última invenção a emergir da Idade da Razão possa vir a ser um «motor da razão» construído sobre o objeto de *software* mais complexo alguma vez criado.[33] Ainda na infância, a IA é já capaz de comparar conceitos, elaborar contra-argumentos e gerar analogias. Está a dar os primeiros passos no sentido da avaliação da verdade e da obtenção de efeitos cinéticos diretos.

O que acontece quando as máquinas atingem um fim de mundo intelectual ou físico? À medida que vão conhecendo e moldando o nosso mundo, é concebível que possam vir a compreender plenamente o contexto da sua criação e talvez ir além do que conhecemos como o nosso mundo. Estamos perante uma transformação magalhânica, desta vez não com a perspetiva de navegar à deriva, mas com o perigo intelectual de enfrentar mistérios que estão para lá dos limites da compreensão humana.

Se a humanidade começar a pressentir a sua possível substituição como principal ator intelectual e físico do planeta, alguns poderão atribuir uma espécie de divindade às próprias máquinas, encorajando ainda mais o fatalismo e a submissão dos humanos. Outros poderão adotar o ponto de

vista oposto: uma espécie de subjetivismo centrado no ser humano que rejeite de modo abrangente a capacidade das máquinas de atingir qualquer grau de verdade objetiva e procure proibir a atividade possibilitada pela IA.

Nenhuma destas posturas mentais permitiria uma evolução desejável ou construtiva do *Homo technicus* – uma espécie humana capaz de, nesta nova era, viver em simbiose com a tecnologia das máquinas.[34] Na realidade, qualquer delas poderia impedir a evolução da nossa espécie. No primeiro cenário, fatalista, poderíamos extinguir-nos. No segundo cenário, de rejeição, ao proibir o desenvolvimento da IA e ao optar pela estagnação, estaríamos a alimentar a esperança de evitar o mesmo destino de extinção – ainda que, dados os riscos existenciais que a nossa espécie enfrenta, incluindo a atual crise diplomática e as mudanças climáticas, essa esperança possa bem estar à partida condenada a ver-se frustrada.

---

28 René Descartes, «Sixth Meditation», *The Philosophical Writings of Descartes*, trad. John Cottingham *et al.*, vol. 2 (Cambridge: Cambridge University Press, 1984), 55.

29 Alfred North Whitehead, *Process and Reality: An Essay in Cosmology,* 2.ª ed. (Nova Iorque: Free Press, 1979), 15.

30 Ver «Debate: Do Language Models Need Sensory Grounding for Meaning and Understanding?», New York University Center for Mind, Brain, and Consciousness, 24 de março de 2023: https://wp.nyu.edu/consciousness/do-large-language-models-need-sensory-grounding-formeaning-and-understanding.

31 Lauren Jackson, «What if A.I. Sentience Is a Question of Degree?», *New York Times*, 12 de abril de 2023.

32 Isto é verdade em algumas teorias da física quântica, em que a observação cria uma mudança objetiva na realidade. Também é verdade no sentido da compreensão humana de uma realidade em que a observação da IA pode criar uma mudança subjetiva; ver Marcelo Gleiser, *The Island of Knowledge* (Nova Iorque: PublicAffairs, 2015), prólogo: «A nossa perceção do que é real evolui com os instrumentos que usamos para sondar a Natureza».

33 Ver os comentários de Ilya Sutskever em Ross Andersen, «Does Sam Altman Know What He's Creating?», *The Atlantic*, 24 de julho de 2023.

34 Para uma exploração mais aprofundada da ideia *Homo Technicus*, ver Henry A. Kissinger, Eric Schmidt, Daniel Huttenlocher, «ChatGPT Heralds an Intellectual Revolution», *Wall Street Journal*, 24 de fevereiro de 2023.

PARTE II

OS QUATRO RAMOS

CAPÍTULO 4

POLÍTICA

Durante anos, correram pelo Novo Mundo rumores a respeito da existência de uma misteriosa e poderosa civilização nas profundezas das selvas do México. A coroa espanhola – a mais ativa e mais bem posicionada entre as rivais europeias – começou a enviar expedições a partir da fortaleza que instalara na ilha de Cuba: um ponto de partida ideal para os seus conquistadores, que tinham sido autorizados a negociar, mas não a conquistar.

Depois de duas expedições falhadas, comandadas por Francisco Hernández de Córdoba e Juan de Grijalva, foi escolhido para liderar a terceira um explorador mais ousado, mais temerário e menos conhecido. Alguns administradores coloniais mostraram-se avessos à escolha, exprimindo o receio de que o novo comandante não tivesse a experiência nem o discernimento necessários para a missão, pelo que esta foi temporariamente cancelada até se encontrar um substituto mais adequado. Mas, quando ninguém o esperava, em fevereiro de 1519, Hernán Cortés, com trinta e quatro anos – determinado a não perder a sua oportunidade de alcançar a glória – escapuliu-se do porto de Santiago a coberto da noite e partiu com onze navios em busca do império oculto.[35]

Os navios espanhóis começaram a ser atentamente vigiados por batedores astecas a partir do momento em que apareceram ao largo da costa de Iucatão. Mas o que esses batedores viram dos visitantes – que afundaram os navios em que tinham chegado, cavalgavam veados de aparência bizarra

e possuíam paus que emitiam relâmpagos – foi inicialmente desconcertante. Os relatórios que enviaram para a capital, Tenochtitlán, chegaram naturalmente às mãos do nono imperador asteca, Montezuma II. Dessa vez, o governante supremo dos Astecas, normalmente decidido e resoluto, não foi nenhuma destas coisas.

Na mitologia asteca, Topiltzin Quetzalcoatl, Senhor dos Toltecas, tinha trazido a civilização e o progresso a esse povo, antecessor dos Astecas. Graças à sua liderança e feitos maravilhosos, foi feito soberano e dado o nome do deus da criação, Quetzalcoatl, e era visto pela maior parte dos Astecas como dotado dos mesmos poderes sobrenaturais da divindade.

Consideremos uma versão dos acontecimentos que nos foi transmitida pelo frade franciscano espanhol Bernardino de Sahagún. Na sua narrativa, Sahagún conta que, segundo a lenda asteca, este homem-deus, que mais tarde viria a cair em desgraça, se aventurou na costa do Golfo onde, à deriva numa pequena canoa de madeira incendiada, jurou regressar dessa mesma direção, atravessando o mar, no então longínquo ano de... 1519.

Na história de Sahagún, o desembarque de Cortés na data prevista para o regresso de Topiltzin incutiu em Montezuma e nos seus conselheiros mais próximos um misto de pavor e respeito, reforçado pela semelhança física de Cortés com as descrições de Topiltzin nos mitos mesoamericanos. Mas uma vez confirmadas através de sucessivos sinais premonitórios – um cometa, um eclipse, o nascimento de um bebé deformado –, as credenciais desta figura semelhante a uma aparição, chegaram prontamente emissários da capital asteca com presentes de ouro e um convite pessoal para Cortés se encontrar com o imperador, pondo em contacto estes poderosos representantes do Novo Mundo e do Velho Mundo.[36]

A história conta que Montezuma se mostrou um anfitrião generoso, cumulando de banquetes e fortuna os seus convidados e oferecendo a Cortés (como era costume na diplomacia asteca) uma das suas filhas em casamento. Mas por trás da faustosa receção, a corte de Montezuma e em breve todo o império estavam fortemente divididos. O irmão mais novo do imperador, Cuitláhuac, chefe das temidas forças de segurança astecas, tinha desconfiado desde o início e pressionava agressivamente para que os estrangeiros fossem mandados embora.

De acordo com relatos de testemunhas oculares, Cortés e os seus associados, alarmados com a precariedade da situação mas encorajados pelas divisões cada vez mais públicas dos anfitriões, fizeram o impensável: capturaram Montezuma e fizeram-no refém. Muitos Astecas vituperaram o ato; outros aplaudiram-no. Durante cerca de um ano, os espanhóis governariam os Astecas por procuração, através de Montezuma, até que as hostilidades explodiram num conflito aberto. Não está claro por que razão e em que medida os anfitriões astecas tornados cativos toleraram este estado de coisas durante tanto tempo, mas a aquiescência do próprio Montezuma – voluntária ou não – ajudou sem dúvida a garantir a submissão temporária do vasto território.

Nenhum presidente, primeiro-ministro, líder supremo, secretário-geral ou monarca dos nossos dias estaria suficientemente preparado para a chegada de forasteiros que, pela oportunidade cronológica e tecnologia avançada, parecessem divinos. Na altura, porém, e durante muito tempo depois, o caso que acabámos de descrever não foi o único. Nos séculos que se seguiram, os estadistas que navegavam nas complexidades das colonizações, consolidações e independências, ou se debruçavam sobre as exigências sociais e políticas de grupos de imigrantes demograficamente poderosos, iriam ter muitas vezes de lidar com modos ainda mais rotineiros e altamente programados de intrusão e transição na liderança.

Agora que a IA chega às nossas proverbiais costas, está, tal como os conquistadores, a provocar sussurros de entusiasmo e de desconfiança. Alguns deixam-se seduzir pelo potencial desta nova fonte de legitimidade para reforçar o seu próprio domínio. Outros esforçam-se por impedir a possibilidade de uma dependência inquebrável ou, pelo menos, a preparar-se para a tarefa de enfrentar as consequências. A IA ativou simultaneamente um impulso para abraçar e para expulsar.

Mas será que as IA vão comportar-se como conquistadores? Estarão os líderes humanos destinados a tornar-se seus intermediários: soberanos sem soberania? Ou talvez as IA divinizadas ressuscitem a outrora omnipresente invocação humana do direito divino, com elas próprias no papel de fazedoras de reis? Ou, em alternativa, serão as IA enxertadas – atabalhoadamente de início, mais perfeitamente depois – nas estruturas

existentes para as complementar e melhorar? Ou serão subjugadas como uma classe migrante indesejável e afastada das alavancas do poder a bem da tranquilidade de uma liderança apreensiva?

Não há dúvida de que a IA vai conferir à nossa espécie um meio até agora inimaginável para fazer avançar o empreendimento da descoberta científica, diminuir os encargos do trabalho e reduzir a miséria da dor. A necessidade ou a conveniência de introduzir a IA nos centros de tomada de decisões políticas é, no entanto, muito menos consensual. Mesmo que fosse possível assegurar resultados extraordinários nesse ambiente, é mais do que natural que hesitemos antes de ceder um tal poder à tecnologia.

No domínio da ciência, os instrumentos ampliaram generosamente o nosso aparelho sensorial. No domínio da exploração, as naves mais diversas protegeram os nossos corpos enquanto nos levavam a fronteiras cada vez mais distantes. Mas o exercício do poder político foi sempre, e por boas razões, um empreendimento antropocêntrico e não tecnocêntrico. Até agora.

A RODA DA HISTÓRIA

Durante grande parte da nossa história, o poder político humano foi considerado um legado divino. A fé e a política estavam intrinsecamente ligadas. Mesmo depois de terem sido separadas pela secularização dos modos de governo modernos, mantiveram certas semelhanças. Como o filósofo inglês G. K. Chesterton observou: «Onde quer que as pessoas não acreditem em qualquer coisa para lá do mundo, adorarão o mundo. Mas, acima de tudo, adorarão a coisa mais forte do mundo.»[37]

Neste aspeto, tanto a religião como a política não só passaram por padrões cíclicos de criação e destruição, como também anteviram a sua recorrência. No teísmo hindu, a evolução natural e social contínua ocorre através de «noites de caos» cíclicas na roda infindável da existência, em que cada período – ou ciclo *yuga* – dura mais de quatro milhões de anos. Muitos hindus acreditam que estamos atualmente a viver a melhor parte da pior era, a *Kali Yuga*, um tempo de escuridão espiritual em que os humanos se julgam falsamente superiores aos deuses.

Também no budismo, ao nível individual, a vida é um arranjo cíclico de morte e reencarnação – uma ideia contada nas escrituras e transcrita na arte. Os monges budistas usam areia colorida para criar mandalas – intrincados diagramas geométricos e cósmicos que podem demorar semanas a fazer – e depois destroem-nas em poucos minutos, para refletir a natureza transitória da vida.[38] O colossal templo budista de Borobudur, na ilha de Java, e o complexo de templos hindu-budistas de Angkor, no Camboja, são vistos por alguns como mandalas arquitetónicas tridimensionais e contam-se ainda hoje entre as maiores estruturas religiosas do mundo.

Todos sabemos como funciona o pêndulo da opinião política. A missão dos líderes esclarecidos é dedicar a vida a manobras capazes de resistir à oposição interna, sobreviver aos inimigos externos e criar novos padrões de relativa paz e estabilidade. Ainda assim, mesmo neste horizonte temporal alargado, todos os líderes políticos agem com o conhecimento de que o seu trabalho não permanecerá por muito tempo após a sua morte, e muito menos após o declínio dos Estados. Tal como os seres humanos individuais, as civilizações acabarão por se desintegrar à medida que as sociedades se desencantam das máximas intelectuais e dos valores subjacentes sobre os quais foram construídas.

Mais ainda, todas as tradições políticas e religiosas reconhecem o potencial de um cataclismo total – seja como fim ou como continuação de ciclos anteriores. Na Bíblia hebraica, o controlo alterna entre Deus e o regente terrestre sancionado, exigindo uma restauração da supervisão divina cada vez que a humanidade, sem a instrução orientadora do Céu, se coloca à beira do desastre. Os budistas procuram a iluminação como saída e libertação de novos ciclos de reencarnação. Os hindus acreditam que a nossa era, a quarta e última *yuga* do ciclo atual, se tornará cada vez mais turbulenta e anárquica até que a calamidade finalmente reinicie o mundo, levando a um regresso à primeira era – a dourada *Satya Yuga*, em que a humanidade é governada por deuses – e deste modo ao início do próximo ciclo. Rezando pela redenção, sacerdotes e monges de diversas crenças prepararam-se para o regresso de um ser superior – uma criança com capacidades sobrenaturais, um redentor designado, o décimo segundo imã – que assinalará o fim ou o princípio da história.

Nos seus estudos matemáticos pioneiros sobre a mudança, Isaac Newton e Gottfried Wilhelm Leibniz, contemporâneos que discordavam em muitas coisas, concordavam, no entanto, que, sob ampliação, qualquer curva parecerá linear, ao ponto de uma descontinuidade iminente quase não ser registada. O advento da IA no domínio do Estado, o domínio mais humano, representaria uma dessas descontinuidades, pressagiando um desenvolvimento possivelmente exponencial das capacidades administrativas humanas, mas ao mesmo tempo um afastamento radical do mundo tradicional do poder e da profecia. A tecnologia torna o progresso inevitável; a política e a fé tornam certa a destruição e a renovação. A IA representa então uma saída dos nossos ciclos familiares ou apenas o início de um ciclo muito mais longo? O princípio do fim, ou apenas o fim do princípio?

POLÍTICA TRADICIONAL

A liderança humana é mais arte do que ciência. Contra probabilidades por vezes insuperáveis, alguns líderes conseguiram resultados excecionais. Na história moderna, entre esses líderes improváveis e bem-sucedidos contam-se Deng Xiaoping na China, Alexander Hamilton nos Estados Unidos e Lee Kuan Yew em Singapura. Os três desencadearam forças sociais subjacentes latentes que escapavam ao controlo de uma qualquer dada pessoa. Deng fundiu o capitalismo com uma antiga burocracia meritocrática que não tinha uma doutrina económica viável; Hamilton permitiu que uma nova filosofia política se espalhasse por uma vasta fronteira em condições de ausência de entidades políticas unificadas. Lee forjou uma ilha de excelência exigindo que ela ascendesse não obstante recursos muito limitados.

Todos estes três líderes combinavam espantosos poderes mentais com uma intensa força de vontade e carisma pessoal. Na linguagem política e noutras formas de retórica persuasiva, os atores humanos tendem a projetar visões para o futuro das suas sociedades que – para usar a expressão de Aristóteles – são uma parte *logos* (lógica), uma parte *ethos* (autoridade

individual) e uma parte *pathos* (ligação emocional). Estas estratégias – em parte morais, em parte psicológicas – são essenciais para a criação e sustentação de identidades culturais unificadas e de sistemas políticos coerentes. Os líderes são, as mais das vezes, contadores de histórias, animando audiências e agitando almas.

Mas a tendência humana para o emocional e o estético pode também ser uma desvantagem. Até os líderes mais sábios, movidos pelo instinto e temperados pela prudência, governam por vezes com base em paixões fugazes. Os governos (tal como as empresas, as igrejas e as famílias, todas elas organizações concebidas e geridas por seres humanos falíveis) são uma combinação imperfeita de tradição herdada e experimentação.

Nas democracias ou nas autocracias, no terceiro mundo ou no primeiro, no passado ou no presente, os seres humanos são praticamente os mesmos. A passagem do tempo ainda não produziu inovações significativas na maneira como nos governamos. Continuamos a utilizar as mesmas instituições que os nossos antepassados já utilizavam há milhares de anos. É certo que ainda nos encontrarmos dependentes de princípios do nosso passado distante não é necessariamente uma coisa má, nem deveria ser particularmente surpreendente, uma vez que a sabedoria dos antigos serviu muitas vezes de inspiração conceptual e de base prática para as nossas sociedades mais bem-sucedidas. Mas essas sociedades podem ser mais a exceção do que a regra. Porque por muitas vezes que figuras excecionais tenham adaptado a tradição histórica para o melhor, um número ainda maior fez que a história pendesse para o pior.

A nossa consistência política pode em parte dever-se à circunstância de sermos humanos – alternadamente leais e caprichosos, humildes e ambiciosos, generosos e egoístas. A nossa volubilidade é mais evidente nos sistemas autocráticos, em que os caprichos de um governante podem prevalecer sobre a coerência demonstrada por outros; a resultante exploração da riqueza nacional e a militarização da justiça são transgressões fáceis de condenar, mas difíceis de eliminar. O nepotismo – que promove a exploração acima mencionada e agrava o militarismo – corrói a fé dos cidadãos que desejam melhorar a condição do Estado e a sua própria situação. Os suficientemente desesperados e corajosos para exigir uma mudança de política – quanto mais uma mudança de regime – devem estar

dispostos a sujeitar-se a uma luta injusta. Os instigadores de revoluções sangrentas são com frequência vilipendiados no imediato, mas – quando bem-sucedidos – glorificados e saudosamente lembrados a longo prazo.

Infelizmente, também a democracia pode ser vulnerável, embora de uma maneira mais subtil, às irracionalidades humanas. Apesar da ausência de um *status* de liderança formalmente herdado, o poder democrático pode ser autoperpetuado. O pressuposto da igualdade, acompanhado pela abstração dos deveres individuais das responsabilidades sociais, pode substituir-se à gradação e à moderação, resultando na absolvição total ou na reprovação total. E, numa época de saturação mediática, é difícil procurar a sabedoria da democracia por trás do ruído. As ideias virais podem assumir uma influência inesperada.

Há problemas que parecem afetar sem exceção todos os sistemas humanos. Avaliar exaustivamente o caminho para uma política perfeita exige o conhecimento de inúmeros fatores esotéricos; com recursos limitados e uma ciência social imprecisa, os resultados divergem muitas vezes da conceção original. Tanto nas autocracias como nas democracias, os políticos – eleitos ou nomeados – podem tomar decisões em parte influenciadas pelo potencial de promoção do seu próprio poder ou lucro. O dinheiro leva aqueles que governam o mundo a comportar-se de uma maneira previsível ou, na ausência de dinheiro, de uma maneira imprevisível.[39]

O reconhecimento e a avaliação do desfasamento entre as nossas expectativas e a realidade têm frequentemente dependido da capacidade de observadores perspicazes, muitas vezes à distância – por exemplo, Thomas Carlyle (sobre a França), Alexis de Tocqueville (sobre a América) e Oswald Spengler (sobre o Ocidente) –, para articular o que, para alguém «de dentro», pode parecer uma falha óbvia mas a que não sabe dar um nome.[40] Acontece, porém, que todos nós, humanos, somos «de dentro» no que respeita à história política coletiva. A falta de variedade nos nossos modos históricos de governação, juntamente com a aparente incapacidade de imaginar alternativas compatíveis com os nossos valores civilizacionais, tem limitado a inovação política. A IA, como elemento externo e perturbador, pode abrir novas possibilidades, mas o custo e as vantagens das alternativas que oferece não estão ainda claros.

O REGRESSO DO REI

É talvez o debate mais antigo da filosofia política: definir ou refutar a viabilidade da sabedoria de uma pessoa contra a sabedoria de muitas. Os filósofos gregos Platão e Aristóteles apresentaram alguns dos primeiros debates concorrentes sobre, entre muitos outros assuntos, a essência de uma governação adequada. Enquanto o primeiro, falando em nome do seu grande antecessor, Sócrates, defendia um «rei-filósofo» singular, dotado de uma sabedoria e vontade aparentemente sobrenaturais, o segundo considerava este ideal interessante em teoria mas impossível na prática, insistindo em que todos os cidadãos participassem em pé de igualdade na administração do Estado.

Durante os dois mil anos seguintes, as ideias de Platão, e as reformulações que se lhes seguiram, seriam repetidamente testadas e consideradas insuficientes na bigorna da aplicação prática. Foi a fórmula de Aristóteles que acabou por se impor, ao princípio devagar, depois rapidamente, quando, no início da era moderna, o filósofo judeu holandês Baruch Spinoza e outros pensadores do Iluminismo desenvolveram filosofias políticas mais seculares. Na China, o estadista, general e filósofo da dinastia Ming, Wang Yangming, cunhou a expressão «unidade de conhecimento e ação» para descrever os dois aspetos do conhecimento que tinham de influenciar a intuição do governante.[41] Os líderes deviam imbuir-se plenamente de filosofia («conhecimento»), desenvolvendo ao mesmo tempo uma competência perfeita em todo o espectro das necessidades burocráticas.

Fosse qual fosse a forma humana em que o ideal socrático-platónico se concretizasse – ditador benigno, déspota esclarecido, um *Übermensch* nietzschiano que personificasse (segundo uma interpretação contemporânea) a «união entre os princípios e visões do mundo apolíneos e dionisíacos» – havia sempre o problema de ter uma figura única no comando, independentemente do intelecto, da experiência ou da visão superiores dessa figura.[42]

Também o pensador islâmico do início da Idade Média, Abu Nasr Muhammad al-Farabi, que levaria as ideias de Platão para o mundo islâmico, propôs adaptações baseadas nas suas observações pessoais da alta

corte da Pérsia. Talvez influenciado pelo seu prodigioso talento para a matemática, concluiu que qualquer indivíduo que possuísse todas as virtudes necessárias procuradas por Platão seria uma enorme anomalia estatística; na prática, portanto, o papel de rei-filósofo devia ser partilhado por duas pessoas, «uma das quais é filósofa e a outra preenche as restantes condições».[43]

Apesar de al-Farabi ter feito tudo para reunir as mentes mais brilhantes do império ao serviço dos vizires persas, só seis séculos após a sua morte o regente muçulmano Bairam Khan (1501-1561) concretizaria a sua visão, orientando Akbar, *o Grande*, e outros imperadores mogóis sob o título de «Vakil»: lugar-tenente de confiança e principal mentor do imperador em todos os assuntos.[44] Mais ou menos pela mesma altura, o diplomata e filósofo italiano Nicolau Maquiavel (1469-1527), convencido não pelas anomalias estatísticas mas pela brutalidade humana, chegaria a conclusões semelhantes às de al-Farabi. Expondo o lado mais obscuro da política e argumentando que o poder só era compatível com o pragmatismo, Maquiavel aconselhava os governantes das cidades-Estado italianas a divorciar a arte da prática e a suspender a ética no exercício da alta política.[45]

Quatro séculos mais tarde, o filósofo germano-americano Leo Strauss ligaria as ideias de Maquiavel às de Platão, concluindo que, para alcançar o modelo ideal do rei-filósofo, a sabedoria teórica e a sabedoria política – ou seja, filósofos e reis – teriam de ser mantidas separadas: «Há um conflito necessário entre filosofia e política se o único elemento [com que] a sociedade contribuiu for necessariamente [apenas] opinião.»[46] Em vez de produzir uma unidade de ambos infundida num único governante, o filósofo governaria o Estado através da *proximidade* do poder. Nesse regime, o filósofo poderia perseguir e aplicar o seu conhecimento acumulado, suficientemente afastado da impureza da política para preservar a pureza do pensamento mas suficientemente próximo para que a sociedade beneficiasse do resultado.

O problema central do rei-filósofo era, e continua a ser, a capacidade limitada de uma única mente humana. Mesmo que fosse possível recolher informação suficiente com a rapidez necessária, os líderes humanos mais experientes possuiriam apenas uma fração da função cognitiva necessária para analisar essa informação e chegar a uma decisão bem fundamentada. O

mundo é demasiado complexo e a intuição humana tem os seus limites. As visões de uma equipa de liderança com duas cabeças eram apenas uma solução parcial para esta dificuldade.

Os seres humanos raramente pensam na administração política como uma capacidade de processar informação. Preferimos elevar a descrição e os objetivos dos nossos sistemas políticos à condição de corporizações de valores e ideologias. Mas, tanto em tempo de paz como de guerra, grande parte da administração de um Estado-nação é uma questão de eficiência no processamento de dados. As falhas no tratamento de informações podem explicar o declínio de muitos sistemas de governo excessivamente centralizados, incluindo o da União Soviética. A dada altura, durante a Guerra Fria, os cientistas soviéticos planearam a criação de um aparelho político cibernético para dotar a liderança comunista de instrumentos tecnológicos destinados a substituir as forças do mercado livre usadas com efeitos significativos pelos seus rivais capitalistas.

O que Platão queria há dois mil anos – e os ciberneticistas soviéticos queriam há quarenta – era um engenheiro de sistemas complexos ou, talvez, uma IA. Mas, antes da invenção da IA, uma autoridade central não podia tomar decisões baseadas no conhecimento local, ou pelo menos não de uma maneira eficiente. Esta ideia foi articulada de maneira mais clara pelo economista e filósofo político do século XX Friedrich Hayek. Só quando refreámos a nossa ânsia de controlo fomos capazes de libertar as forças invisíveis que impulsionam a atribuição de talento, riqueza e ideias em modos que há muito nos escapavam.

Uma das principais desvantagens dos sistemas centralizados, quando a informação ainda tinha de viajar por terra e não como ondas pelo ar, era o tempo de atraso na comunicação. Alguns supuseram que, no século XXI, a velocidade quase instantânea da informação permitiria que os sistemas centralizados se equiparassem aos difusos. O que ainda não aconteceu. O grande obstáculo não tem sido a rapidez, mas a complexidade.

Os atributos das máquinas discutidos no capítulo anterior tornam agora possível a uma IA centralizada competir com processadores de informação descentralizados e potencialmente superá-los – com a sua velocidade a reduzir ainda mais a tensão latente entre teoria e realidade e a sua escala e

resolução a garantirem um âmbito e uma precisão abrangentes.[47] A resposta democrática de Aristóteles pode, de facto, ter sido mais ética do que a de Platão, mas ganhou a batalha da história porque foi a mais eficiente.

Isto não significa que as capacidades de processamento de informação da IA tenham de perturbar a própria democracia, mas é um facto que a possibilidade de uma centralização eficiente está preparada para remodelar os canais através dos quais a democracia é implementada. Poderá resultar na instanciação de uma bolsa de valores democrática, um mercado único de ideias, avaliadas e valorizadas a uma velocidade incrível. As IA podem ser capazes de incorporar as perspetivas de milhares de milhões ou mesmo de biliões de futuros humanos, estimando com precisão os benefícios e os danos atuais e distantes no tempo de – por exemplo – um progresso tecnológico disruptivo.

Juntas, as imaginações coletivas destas mentes poderiam então ser concentradas numa única força, que igualaria os métodos aristotélicos e potencialmente se expandiria para lá deles. Se fosse possível conceber uma única mente para representar a inteligência coletiva e os valores de uma política democrática – um imenso filósofo para estar ao lado do rei (eleito) – então a visão de duas cabeças de al-Farabi e Strauss poderia ser realizada.

Eventualmente, a IA poderá acabar com o fosso entre as ciências técnicas e as ciências sociais – sendo a ciência política a principal destas últimas. As leis da natureza humana poderão tornar-se tão previsíveis, e a política tão fiável, como as leis da física o são atualmente para nós.

Um processamento de informação sem precedentes permitirá uma centralização verdadeiramente eficiente da política pelas IA. Poder-se-ia esperar que isto reforçasse a perceção de controlo por parte das elites. No entanto, a opacidade destes sistemas – e a noção de que o seu funcionamento pode ser otimizado na ausência de interferência humana – irá agir na direção oposta. É possível que, com o tempo e a experiência, o controlo humano possa vir a parecer menos uma necessidade do que um fardo. Mesmo que, no início, os líderes europeus do século XVIII se tenham sentido aterrorizados por terem de entregar o controlo às forças invisíveis do interesse próprio humano, os líderes políticos do século XXI poderão ter de se humilhar mais uma vez perante um sistema que incorpora a sabedoria das massas de uma maneira inteiramente nova.

GOVERNAR PELA RAZÃO

Se e quando a IA começar não só a processar informação com objetivos políticos mas também a tomar decisões políticas, novas questões vão pôr novos problemas à sabedoria política convencional. E a ciência política nem sequer oferece palavras orientadoras para uma tal mudança. Como poderá um observador avaliar a «superioridade» de uma decisão estratégica tomada por uma IA na ausência de interpretação da lógica através da qual a IA a tomou? O resultado, claro, só poderia ser métrico. Mas a falta de um registo definidor dos princípios de ação faria que algo muito significativo se perdesse, sobretudo para os historiadores.

Em muitos casos futuros, os humanos poderão discordar dos planos gerados pela IA não por esses planos serem insensatos mas porque as razões que fundamentam uma decisão da IA se situam para lá da sua compreensão imediata. Pode ser particularmente natural querer impedir ou adiar um resultado em que os humanos perdem não só o controlo temporal sobre os processos de tomada de decisão das máquinas (a capacidade de intervir), mas também a compreensão lógica desses processos (a capacidade de interpretar), mesmo depois do facto consumado.

Nas situações em que a liderança humana e da IA divirjam ou discordem em matérias de política de Estado, que julgamento deve prevalecer? Será que a resposta muda caso se torne evidente que a orientação da IA, mais do que a opinião humana, tem efetivamente em conta a concretização dos objetivos humanos no futuro distante, sendo por conseguinte mais vantajosa?

A parte utilitária da intuição humana estaria inclinada a aceitar uma avaliação orientadora a longo prazo gerada por uma IA, sobretudo se essa IA fosse capaz de explicar a fundamentação das suas decisões. Mas, mesmo assim, poderiam surgir situações em que os humanos protestariam contra uma política que, apesar de assegurar a longevidade da espécie e a de pessoas ainda não nascidas, seria prejudicial para os que vivem aqui e agora. Do mesmo modo, a IA poderia aceitar resultados racionais, como a solução de conflitos militares, que fossem politicamente inaceitáveis para todas as partes envolvidas. As probabilidades de rebelião declarada seriam muito elevadas. Mesmo que o conceito de governação pelas máquinas fosse

aceite, e mesmo que a lógica das máquinas fosse explicada, e mesmo que as decisões das máquinas fossem racionais, benéficas e superiores em diversos graus, poderíamos continuar a ser ingovernáveis. Entre os humanos, só as políticas que incorporem um elemento do intangível são sustentáveis. Como Tolstoi escreveu, «Se admitirmos que a vida humana pode ser regida pela razão, a própria possibilidade de vida é destruída.»[48]

Esses elementos não racionais da experiência humana – a história, acima de tudo, mas também a estética, o carisma e a ressonância emocional – podem, em certo sentido, impedir resultados ótimos, mas são também fundamentais para os nossos grupos políticos. O governo apenas pela razão pode dissolver nações. As IA demasiado racionais e os seus parceiros humanos podem rapidamente perder o controlo ou causar a desintegração das estruturas de poder em que cooperam.

Por outro lado, a IA pode ser mais valiosa e mais necessária precisamente onde a sua lógica parece estranha, contraintuitiva ou simplesmente errada. Embora as IA possam ser utilizadas para acelerar a resolução de problemas no sentido de soluções conhecidas – expandindo as opções humanas graças a uma velocidade que evita os custos políticos do atraso –, talvez a sua principal utilização seja pensar no que não temos capacidade para ponderar e encontrar soluções totalmente novas. Na realidade, pode ter sido um dos principais objetivos da sua criação.

Mas também esta atitude acarreta riscos, proporcionais à abertura a oportunidades que oferece. Não deixa aos humanos qualquer base para corrigir ou ignorar as decisões potencialmente inaceitáveis e anteriormente inimagináveis de uma IA ininterpretável. O impulso para não lhe estorvar a ação cresceria, exacerbado pela superioridade potencialmente evidente da sua governação quando comparada com a governação humana do passado.[49] Um governo IA poderia produzir resultados verdadeiramente imbatíveis. Se assim fosse, interromper a sua utilização ou circunscrever o seu âmbito pareceria ilógico – sobretudo no contexto da competição geopolítica, em que a abstenção de utilização pareceria assegurar desvantagem.

Do mesmo modo, os líderes humanos que se habituassem a resultados superiores proporcionados pela IA ficariam dependentes dela para afirmar a sua própria legitimidade. Além disso, a IA pode também desenvolver os

seus próprios preconceitos: se um líder humano leal a um parceiro de IA quisesse manter-se no cargo para lá do seu mandato, iria a IA intervir para impedir essa violação do protocolo?

PROMETEU

Ao longo da História, os líderes que afirmavam saber melhor do que próprio povo o que mais convinha a esse mesmo povo foram rapidamente desmentidos pela realidade.[50] Friedrich Hayek avisou que o planeamento central ou modos semelhantes de governação – incluindo os ainda não conceptualizados no seu tempo – teriam como corolário inevitável a proibição da dissidência.[51] Sem, ou mesmo com, informação exata, o planeamento pode tornar-se um forte argumento para dar prioridade ao coletivo sobre o individual – uma busca do utilitarismo sem obstruções. O que não é inerentemente mau; mas, a partir de um certo ponto, esse governo pode atingir a velocidade de escape, saindo do domínio da governação normal e tornando-se eterno em ubiquidade – com os cidadãos forçados a ser livres, coagidos para seu próprio bem. Uma administração alimentada pela IA poderia, nas palavras de um neuropsicólogo, «afirmar que sabe o que o seu povo realmente quer e o que irá torná-lo verdadeiramente feliz (...) na melhor hipótese [usado] para justificar o paternalismo, e na pior o totalitarismo».[52]

Desde que passámos a conhecer-nos melhor do que o nosso rei nos conhece, o liberalismo tem sido uma força de contenção. Agora, como já foi sugerido, a IA «vai dizer-nos quem somos antes de nós próprios [o sabermos]», dotando deste modo os totalitaristas não apenas de uma ferramenta operacional mas também de uma «arma filosófica».[53] A IA poderia então contradizer as afirmações de Immanuel Kant:

> Ninguém tem o direito de me obrigar a ser feliz da maneira peculiar como entende o bem-estar alheio; mas todos têm o direito de procurar a sua própria felicidade da maneira que melhor lhes pareça, desde que não infrinja a liberdade de terceiros na demanda de um objetivo igual

> para si mesmos sempre que a Liberdade desses terceiros seja consistente com o Direito à Liberdade de todos de acordo com as leis universais possíveis.[54]

No processo de construção da IA, as máquinas aprendem que os erros são «errados». Isto só é útil na medida em que corrige as assunções erróneas e as implementações menos do que ótimas dos humanos. Mas outra causa de erro nos nossos sistemas humanos é o livre arbítrio individual – somos livres de fazer a escolha «errada». Se um sistema de IA decidir eliminar esses erros, tem duas opções: remover-nos, ou retirar-nos o livre arbítrio. Se o livre arbítrio for, por defeito, considerado não uma característica da inteligência mas um erro, pode também ser cada vez mais visto como uma barreira que impede a IA de atingir os seus objetivos.

Os seres humanos habituaram-se a apreciar e a desprezar o facto de a força de vontade ser o que faz mover o mundo. A história foi escrita de modo a mostrar os seres humanos como os seus principais autores; nela, os líderes individuais exerceram a autoridade para garantir que os impérios se erguem e caem à sua ordem ou que os grandes conquistadores podem impor estabilidade suficiente para domar o caos. O descontrolo total pode ser rejeitado, mesmo sabendo-se que maximiza a capacidade da IA para agir no nosso interesse (e no das gerações futuras).

O equilíbrio entre controlo e descontrolo pode ser influenciado pelo problema do tempo discutido no capítulo anterior, com as perceções da humanidade e da IA a operar em escalas temporais diferentes. É pouco provável que a IA se sinta angustiada com a velocidade a que atua; verá como natural a taxa de mudança que cria. Para os humanos, pelo contrário, as vantagens acrescidas da IA poderiam ser tão espetaculares – e tão estranhamente rápidas – que, em pouco tempo, a nossa espécie acabasse por sentir que já bebera dela o máximo que era capaz de tolerar e que continuar a beber seria mais desorientador do que agradável, não valendo, por conseguinte, o preço a pagar.

O que não significa necessariamente o fim da mudança. Significa apenas que, para trazerem consigo sociedades intactas, as vantagens que é possível recolher da IA teriam de ser incorporadas nas instituições humanas pouco a pouco, ou pelo menos mais devagar do que os maximizadores de distantes bens futuros talvez gostassem.

Muitas pessoas conhecem, por exemplo, a história de Prometeu, o titã que roubou fogo aos deuses para o oferecer aos humanos.[55] Menos saberão que, ao fazê-lo, estava perfeitamente consciente do cruel castigo que Zeus lhe imporia: ser acorrentado a um rochedo por toda a eternidade enquanto uma águia iria diariamente devorar-lhe o fígado (que se regeneraria durante a noite). O que o mito deixa sem resposta, no entanto, é por que motivo havia Prometeu de decidir sacrificar-se sabendo antecipadamente, como também sabia, todos os horrores que a humanidade iria desencadear com a sua oferta. Talvez tenha feito a escolha que fez, e seja considerado um herói em vez de um vilão, porque a humanidade também usou o fogo para coisas muito boas.

Restringida a parcerias com os humanos, em vez de largada à solta no meio deles para os dominar, a IA pode ajudá-los na governação de maneiras obscuras reminiscentes e reflexas de um género semelhante de antevisão inacessível – sendo precisamente «antevisão» o significado do nome Prometeu em grego. Mas só podemos aceitá-lo como herói se ele exercer essa antevisão ao serviço da humanidade, uma humanidade que conserve uma verdadeira capacidade individual de ação e coerência política, mesmo que através de alguma irracionalidade.

UM NOVO ATO

Quando examinamos o registo histórico, talvez o que nos pareça mais notável não seja a quantidade de mudança que vemos na política mas, pelo contrário, a ausência de mudança. Persistem hoje os mesmos arquétipos de liderança que têm existido desde há milhares de anos: o trágico príncipe-herói, o conselheiro velhaco, o fiel lugar-tenente, o bobo da corte, o mestre-bonecreiro a puxar os cordelinhos na sombra, o mercenário indigno de confiança. Se a política é um teatro, estas personagens familiares ajudam-nos a explicar o de outro modo inexplicável e dar vida ao de outro modo arcano.

Antes de Catarina ser grande, Jaroslau sábio, Ivã terrível ou Suleimão magnífico, nenhum deles era nenhuma destas coisas. E nós admiramos estes líderes em parte porque a História deu testemunho das respetivas evoluções individuais. Vemos com particular admiração aqueles que, nascidos em circunstâncias desfavoráveis, souberam à custa de pura convicção libertar-se das grilhetas da adversidade e ascender ao topo das suas sociedades. O rei mameluco Iltutmish, vendido como escravo pelos próprios irmãos, que lhe invejavam a beleza física e a inteligência, iria servir mercadores de escravos usbeques e afegãos em Bukhara e Gázni antes de ser comprado por um comandante escravo gúrida num mercado de Deli. Volvidas apenas duas décadas, ele, o escravo de um escravo, subiria nas fileiras até ao serviço do sultão, fundando uma «dinastia de escravos» no topo do reino dos seus antigos senhores.

Claro que a queda do poder pode ser tão inesperada como a ascensão. Nenhum líder tem garantida imunidade contra ser derrubado por uma revolução ou enforcado por conspiração. Seja pelo suicídio como Nero, num duelo como Hamilton ou o assassínio como Gandhi, a grandeza não pode escapar à sepultura.

Como líderes na História, estas figuras partilham o mesmo interesse na ordem e na segurança básicas. Mas como membros da família do homem não estão, como não está qualquer um de nós, livres de inveja, desconfiança recíproca e rivalidade entre imãos.[56] Encontramos as mesmas lutas tanto em nossas casas como nas casas governantes, em gerações de pessoas comuns e em dinastias reais. Podemos compreender com empatia a dificuldade de ultrapassar a inveja e forjar alianças com antigos inimigos – como fez a imperatriz-viúva Cixi, uma ex-concubina que acabaria por governar a China –, tal como podemos desprezar a traição do general Mir Jafar, cuja deserção na batalha de Plassey entregou efetivamente o controlo da Índia aos conquistadores britânicos a troco do título mais do que honorário de «Primeiro Nababo de Bengala». Se os líderes fossem demasiado parecidos com deuses, demasiado perfeitos, ou demasiado competentes, não sentiriam ansiedade, desespero, amor ou inveja. As familiares emoções e vulnerabilidades que todos os humanos conhecem também fazem girar a roda das História.

A política humana é ao mesmo tempo louvável e condenável devido à sua proximidade espiritual da nossa política individual. Apesar de ter conquistado a Europa, Napoleão não foi capaz de conquistar o coração de Joséphine; a rivalidade entre John Adams e Thomas Jefferson parece-nos tão familiar como as nossas quezílias de irmãos. O *Guerra e Paz* de Tolstoi é tanto a respeito dos principais acontecimentos da História como das vidas individuais que habitam esses acontecimentos.[57] Os homens que no século XV seguiram a jovem Joana d'Arc para a guerra não eram muito diferentes dos que em 324 a.C. se amotinaram em Ópis contra Alexandre, *o Grande*, depois de ele ter conquistado a maior parte do mundo conhecido. Alternando entre o político e o pessoal, a realidade funde-se na ficção, as crónicas históricas em poemas épicos.

Uma política mecânica não teria esta tensão narrativa. Sem a purga de rivais nem a reaproximação a antigos inimigos, a súbita ascensão ao poder e a rápida queda em desgraça, a atividade dos estadistas tornar-se-ia menos relatável, menos emocionante, se não mesmo enfadonha. Sem distinção entre tragédia e comédia, a corte imperial ficaria despojada de drama e intriga.

Alternativamente, a emergência da IA como característica nova poderia representar o incitamento a um novo ato na política humana. Poderia alterar a natureza e a dinâmica dos conhecidos arquétipos humanos. Mas algumas coisas não mudariam. O tempo finito da vida humana garante os nossos arcos narrativos, as nossas ascensões e quedas. A nossa natureza social evoluída dita os nossos amores, ambições e valores morais.

É esta mesma característica imperfeita da nossa política que devíamos ter como objetivo preservar e posteriormente combinar com a perfeição dos sistemas de IA para torná-los parceiros complementares. A IA, faltando-lhe as inconsistências da alma humana, será imune a constrangimentos; para o melhor e para o pior, a nossa combatividade tem funcionado como um travão do nosso potencial, mas também como uma fiável restrição do mal que somos capazes de fazer.

O governo humano dos tempos atuais assenta na nossa longa experiência com a contingência histórica. O valor da IA na governação, até agora adormecido, reside no seu conhecimento potencialmente perfeito. Os líderes humanos de hoje deviam preparar-se para ser os primeiros numa linha de soberanos que vão enfrentar o desafio de procurar um equilíbrio

entre maximizar as vantagens – e, em alguns casos, a necessidade – da entrada da IA na governação sem chegarem ao ponto de sucumbir a uma dependência total, descobrindo em vez disso a síntese adequada entre os extremos de despotismo e anarquia, fundindo num todo a vontade dos humanos, o conhecimento das máquinas e a sabedoria da História.

---

35 Ver Salvador de Madariaga, *Hernán Cortés: Conqueror of Mexico* (Nova Iorque: MacMillan, 1941), 99.

36 A narrativa do encontro entre Cortés e Montezuma continua a ser um episódio histórico controverso. Para relatos espanhóis, ver «Historia verdadera de la conquista de la Nueva España) de Bernal Díaz del Castillo, finais do século XVI; Hernan Cortés, «Cartas de Relación» entre 1519 e 1526; «The Florentine Codex» (Historia general de las cosas de Nueva España), finais do século XVI; Carta do núncio apostólico Bernardino de Sagahun, 1524. Para uma interpretação alternativa que questiona o relato histórico dos espanhóis, ver Camilla Townsend, «Burying the White Gods: New Perspetives on the Conquest of Mexico», *The American Historical Review*, vol. 108, n.º 3 (junho de 2003), pp. 659-87) Oxford University Press; «Anales de Tlatelolc», século XVI, e Diego Duran e Alfredo Chavero, Apendice-Explicacion del Codice Geroglifico de Mr. Aubin de Historia de las Indias de la Nueva España y Islas de Tierra Firme, vol. II, 1880, 71.

37 G. K. Chesterton, «Lecture 65: Christendom in Dublin», em *Collected Works*, vol. XX (São Francisco: Ignatius Press, 2002).

38 Ver em especial o Kālacakra.

39 Ver comentários de Chamath Palihapitiya na Stanford Graduate School of Business, 13 de novembro de 2017, https://www.youtube.com/watch?v=PMotykw0SIk.

40 Ver Alexis de Tocqueville, *Democracy in America*, trad. Henry Reeve, Esq., em dois volumes (Londres: Saunders and Otley, 1835; Nova Iorque: J. & H. G. Langley, 1840).

41 Wang Yangming, *Instructions for Practical Living or Record of Transmitting the Mind (Chuanxilu),* postumamente compilado pelos discípulos com base nos seus ensinamentos e em discussões posteriores à sua morte em 1529.

42 Abir Taha, «Nietzsche's Superman», *Artkos* (UK: Artkos Media, 2013), 93.

43 Al-Farabi, *Al-Farabi On the Perfect State*, trad. Richard Walzer (Oxford, Clarendon Press, 1985), 253.

44 Ver T. C. A. Raghavan, *Attendant Lords: Bairam Khan and Abdur Rahim* (Uttar Pradesh: HarperCollins, 2017).

45 Ver Nicolau Maquiavel, *The Prince*, trad. Tim Parks (Londres: Penguin Classics, 2009).

46 Leo Strauss, *What is Political Philosophy?* (Chicago: University of Chicago Press, 1959).

47 Ver Johan Norberg, *The Capitalist Manifesto* (Londres: Atlantic Books, 2023).

48 Lev Tolstoi, *War and Peace*, trad. Louise e Aylmer Maude (Chicago: Encyclopedia Britannica, 1952), 646.

49 Ver Friedrich Hayek, «The Use of Knowledge in Society», *The American Economic Review*, setembro de 1945, e Thomas Sowell, *Knowledge and Decisions* (Nova Iorque: Basic Books, 1996), que elabora sobre o pensamento de Hayek.

50 Ver Hannah Arendt, *The Origins of Totalitarianism* (Nova Iorque: Harcourt, Brace, Jovanovich, 1951).

51 Ver Friedrich Hayek, *The Road to Serfdom* (Chicago: University of Chicago Press, 1944).

52 Simon McCarthy-Jones, «Artificial Intelligence is a totalitarian's dream – here's how to take power back», *The Conversation*, 12 de agosto de 2020, https://theconversation.com/artificial-intelligence-is-a-totalitarians-dream-heres-how-to-take-power-back143722.

53 *Ibid.*; ver Yuval Noah Harari, *Homo Deus: A Brief History of Tomorrow* (Nova Iorque: Harper, 2017).

54 Immanuel Kant, *Kant's Principles of Politics*, trad. W. Hastie (Edimburgo: T. & T. Clark, 1891), 36.

55 Hesíodo, *The Theogony* (Nova Iorque: Start Publishing, 2017); Ésquilo, *Prometheus Bound*, trad. Deborah H. Roberts (Indianapolis: Hackett Publishing Company, 2012).

56 Inspirado em comentários feitos por Lawrence H. Summers no Harvard College China Forum, 17 de abril de 2022.

57 Ver Lev Tolstoi, *War and Peace*, trad. Louise and Aylmer Maude (Chicago: Encyclopedia Britannica, 1952).

CAPÍTULO 5

SEGURANÇA

DA RECALIBRAÇÃO da estratégia militar até à reconstituição da diplomacia, a IA tornar-se-á um fator determinante da ordem no mundo. Imune a medo e a favores, introduz uma nova possibilidade de objetividade na tomada de decisões estratégicas. Mas essa objetividade, aproveitada tanto pelo belicoso como pelo pacificador, deve preservar a subjetividade da sabedoria humana, essencial para o exercício responsável da força. Mais uma revelação da nossa condição atual do que uma nova descoberta do desconhecido, a IA na guerra irá destacar as melhores e piores expressões da nossa humanidade. Mesmo antes da implementação de progressos sérios, vai nascer uma compreensão de como a IA pode ser, ao mesmo tempo, o meio para um conflito e uma maneira de lhe pôr fim.

A antiga luta da humanidade por constituir-se em arranjos cada vez mais complexos de modo que nenhum Estado fique em posição de dominar completamente qualquer outro chegou à condição de uma lei contínua, ininterrupta, da natureza. Num mundo onde os principais atores continuam a ser humanos – mesmo que equipados com IA para os informar e aconselhar – devíamos desfrutar ainda de um certo grau de estabilidade baseada em normas de conduta partilhadas entre os atores envolvidos, sujeitas às afinações e aos ajustamentos do tempo.

Mas se a IA emergir como um conjunto de entidades políticas, diplomáticas e militares praticamente independentes, seremos forçados a trocar o antigo equilíbrio de poder por um novo e desconhecido

desequilíbrio. O concerto internacional dos Estados-nações – um frágil e mutável equilíbrio conseguido nos últimos séculos – tem-se mantido em parte devido à igualdade inerente dos intervenientes. Um mundo de grande assimetria – por exemplo, se alguns Estados adotassem a IA como liderança política mais rapidamente do que outros – seria muito menos previsível. Nos casos em que alguns humanos se confrontassem militar ou diplomaticamente com um Estado altamente capacitado em IA, ou com a própria IA, os humanos poderiam parecer incapazes de sobreviver, quanto mais de competir. Uma tal ordem intermédia poderia assistir a uma implosão das sociedades e a uma explosão incontrolável de conflitos externos.

Há muito que, além de um mínimo de segurança, os motivos que levam os humanos a lutar são a busca do triunfo ou a defesa da honra. Mas as máquinas que – pelo menos por enquanto – não têm qualquer conceção real de triunfo ou de honra poderiam provocar um tipo de guerra único na experiência humana. A questão é: que escolhas fariam as IA? Nunca iriam para a guerra, optando em vez disso (por exemplo) por transferências de território imediatas, cuidadosamente estabelecidas com base em complicados cálculos de vantagens estratégicas relativas? Ou – privilegiando um resultado em detrimento de vidas individuais – permitiriam que os conflitos evoluíssem para guerras de desgaste humano? Num cenário, a nossa espécie poderia emergir transformada ao ponto de evitar totalmente a brutalidade da conduta humana. No outro, ficaríamos subjugados ao ponto de nos fazer voltar a um passado bárbaro.

Não haverá um terceiro caminho? Até à data, nem as pacientes aberturas da diplomacia nem os chocantes horrores da guerra conseguiram codificar permanentemente no ser humano – para não falar de outras espécies – uma aversão à destruição. Será mesmo assim possível que, graças à era da IA, se materializem as condições que permitam finalmente o estabelecimento de uma paz perpétua?

ESPIONAGEM E SABOTAGEM

Ao mesmo tempo que procuram modos de manter a tecnologia segura, as nações estão empenhadas em «ganhar a corrida à IA».[58] Esta reação é em parte compreensível. A cultura, a história, a comunicação e a perceção conspiraram para criar entre os principais Estados atuais uma situação diplomática que promove insegurança e desconfiança de todos os lados. A IA, enquanto característica dominante do futuro global, constitui um ponto de deflagração num conjunto já de si volátil, em que cada uma das partes acredita que uma vantagem tática cada vez maior pode ser decisiva para seu benefício a longo prazo.

Se, motivada pelo instinto de autopreservação, cada sociedade humana quiser maximizar a sua posição unilateral, estarão criadas as condições para uma competição psicológica entre forças militares e agências de informação rivais diferente de tudo o que a humanidade viu até agora. Atualmente, nos anos, meses, semanas e dias que antecedem a chegada da primeira superinteligência, espera-nos um dilema de segurança de natureza existencial. O primeiro desejo lógico de qualquer ator humano que venha a entrar na posse de uma tal capacidade poderá ser tentar garantir a continuação de uma existência singular. Qualquer um desses atores pode também assumir, por defeito, que o seu rival, com as mesmas incertezas e enfrentando os mesmos riscos, estaria a ponderar uma ação semelhante.

Mesmo que uma nação dominante não chegue à guerra, uma IA superinteligente pode subverter, minar e bloquear um programa concorrente. Por exemplo, a IA promete reforçar os vírus informáticos convencionais com uma potência sem precedentes e disfarçá-los com igual eficácia. Tal como o *computer worm* Stuxnet – a arma cibernética que se diz ter destruído um quinto das centrifugadoras de urânio de Teerão antes de ser detetada – um agente de IA poderia sabotar o progresso de um rival de modo a ocultar a sua presença, conduzindo assim os cientistas inimigos por caminhos de investigação fúteis.[59]

Com a sua capacidade única de manipular *en masse* as fraquezas da psicologia humana, uma IA poderia também sequestrar os meios de comunicação de uma nação rival, produzindo um dilúvio de desinformação sintética tão alarmante que inspiraria uma oposição maciça ao progresso das capacidades de IA nesse país. Ou poderia visar o cientista-chefe de IA de

outra nação com comunicações direcionadas e personalizadas que tornassem a perceção de sofrimento pessoal dessa pessoa ao ponto de degradar a sua capacidade de liderança eficaz.

Avaliar o estado da concorrência será ainda mais difícil. Os maiores modelos de IA já estão a ser treinados em redes seguras desligadas do resto da Internet. Alguns executivos acreditam que o próprio desenvolvimento da IA irá, mais cedo ou mais tarde, migrar para *bunkers* impenetráveis cujos supercomputadores serão alimentados por reatores nucleares.[60] Neste preciso momento estão a ser construídos centros de dados no fundo do oceano.[61] Em breve poderão vir a ser colocados em órbita à volta da Terra. As empresas ou as nações poderão cada vez mais entrar em *blackout*, deixando de publicar resultados da investigação sobre IA, não só para limitar a capacidade de ação de agentes maliciosos (como é publicamente afirmado), mas também para ocultar o seu próprio ritmo de construção (como é privadamente pretendido). Para distorcer a imagem real dos seus progressos, outros poderão mesmo tentar publicar investigação deliberadamente falsa, com a IA a ajudar na criação de fabricações convincentes.

Existe um precedente para este tipo de subterfúgio científico: em 1942, o físico soviético Georgy Flyorov inferiu corretamente que os EUA estavam a construir uma bomba nuclear depois de ter reparado que os americanos e os britânicos tinham de súbito deixado de publicar artigos científicos sobre fissão atómica.[62] Hoje, porém, uma tal competição seria ainda mais imprevisível, dada a complexidade e a ambiguidade de medir o progresso em direção a algo tão abstrato como a inteligência. Embora alguns considerem que a «vantagem» é proporcional à dimensão dos modelos de IA que possuem, um modelo maior não é necessariamente superior em todos os contextos e pode nem sempre prevalecer sobre modelos mais pequenos e inferiores utilizados em grande escala. As máquinas de IA mais pequenas e mais especializadas podem funcionar como um enxame de drones contra um porta-aviões – incapaz de destruí-lo, mas suficiente para neutralizar grande parte da sua capacidade.

Para alguns atores**,** a obtenção de uma dada capacidade seria o indicador de uma vantagem global. O problema com esta linha de pensamento é que a IA se refere apenas a um processo de aprendizagem automática que está não integrado numa única tecnologia mas num vasto espectro de tecnologias.

Assim, a capacidade numa área pode ser determinada por fatores totalmente diferentes de capacidade noutra. Neste sentido, qualquer «vantagem» tal como é habitualmente calculada pode ser ilusória.

Além disso, como ficou demonstrado pela explosão exponencial e imprevista da capacidade de IA nos últimos anos, a trajetória do progresso não é linear nem previsível. Olhando para o futuro, os especialistas continuam a discordar sobre o desenvolvimento da superinteligência: será apenas uma questão de escalonamento e de implementação das infraestruturas de aprendizagem existentes, ou a superinteligência vai exigir inovações criativas e científicas adicionais?[63] É possível que a transição da inteligência restrita para a inteligência geral e depois para a superinteligência aconteça sem sinais óbvios de evolução – sobretudo se os humanos não desenvolverem uma conceção uniforme do que procurar. Mesmo que se possa dizer que um ator está um determinado número de anos ou meses à frente de outro, um súbito avanço técnico ou teórico numa área-chave num momento crítico poderia inverter as posições de todos os intervenientes.

Num mundo assim, onde nenhum líder poderia confiar na sua inteligência mais sólida, nos seus instintos mais primários, ou até na base da própria realidade, nenhum verdadeiro candidato à superinteligência pode ser censurado por agir a partir de uma posição de máxima paranoia e suspeita. Os líderes já estão sem dúvida a tomar decisões partindo do pressuposto de que os seus esforços estão a ser vigiados ou que contêm distorções criadas por influência maligna. Assumindo à partida o pior dos cenários, o cálculo estratégico de qualquer ator na fronteira seria dar prioridade à velocidade e ao secretismo em detrimento da segurança. Com a perceção humana incapaz de detetar as tentativas de supressão permitidas pela IA, ou defender-se delas, os líderes humanos poderiam deixar-se dominar pelo medo de não haver um segundo lugar. Sob pressão, poderiam acelerar prematuramente a implantação da IA como modo de dissuasão contra perturbações externas.

De momento, encontramo-nos ainda do lado de cá do véu da ignorância. Os vencedores finais (se é que é possível definir vitória) da competição pelo domínio da IA são por enquanto desconhecidos. Todas as empresas que procuram chegar à primeira linha são potenciais rivais. Esta incerteza vai produzir instabilidade.

Tradicionalmente, após a emergência de uma nova potência tem sido necessária uma contenda sangrenta antes de os rivais conseguirem definir um novo *status quo* que seja, pelo menos, minimamente aceitável para todas as partes. No entanto, num mundo de armas nucleares complicado por uma reconfiguração tão rápida operada pela IA, pode não haver oportunidade de validar um novo consenso com recurso aos princípios estabelecidos e aos instrumentos da guerra.

No caso de vir a cristalizar-se a identidade de um vencedor, a mera competição poderá transformar-se em conflitos impulsionados pelo desespero e pelo medo. Nessas circunstâncias, a certeza poderia ser ainda mais desestabilizadora do que a precaução. Em conjunto, a dissuasão perfeita com velocidade ilimitada e precisão máxima traduz-se em vitória total. A busca histórica do monopólio da força tem normalmente provocado uma consequente expansão da montagem de projetos obscuros pelos «outros». Quando essas capacidades são vistas no contexto de um mundo precariamente equilibrado, alguns Estados podem considerar o advento da IA suficientemente ameaçador para exigir uma resposta nuclear. Será que, tendo evitado as guerras convencionais, a IA nos vai arrastar para o holocausto atómico?

Em anteriores corridas aos armamentos, o instinto evolutivo misturado com a competição entre inventores permitiu a criação e o funcionamento de instrumentos destinados a preservar as sociedades nacionais. Também aqui os sistemas defensivos de IA poderiam neutralizar os ataques dos adversários, por exemplo, atualizando o *software* e outros sistemas atualmente vulneráveis ou, se forem encarregados da vigilância de programas concorrentes, servindo de sistema de alerta precoce. Neste caso, porém, novas ameaças – uma arma biológica ativada por IA, uma súbita proliferação nuclear ou até uma IA desalinhada – podem surgir tão rápida e silenciosamente e causar uma tal devastação que impeça qualquer género de resposta.

Embora possamos não ter ainda chegado a esse ponto, é preciso tomar medidas prévias para gerir a concorrência existencial da era da IA e os riscos que lhe estão associados. Um ator determinado ou desesperado, com uma força secundária mas considerável, estará vigilante na procura de IA superiores. Se esse ator secundário sentir – mesmo que erradamente – que o outro dominante está à beira de conseguir uma capacidade totalizadora,

pode lançar ataques eletrónicos ou físicos preventivos, desencadeando um ciclo de escalada e retaliação inimaginável e a possibilidade de destruição mútua em espiral.

Um vislumbre de esperança – pelo menos por hoje – reside na nossa incerteza. A ambiguidade pode ser um terreno fértil para cultivar o diálogo. Sem saber ao certo de que modo uma nação ou um grupo pode ser beneficiado ou prejudicado pelas decisões de IA tomadas hoje, os líderes globais têm uma janela de oportunidade para se envolverem em discussões com base na sobrevivência coletiva da humanidade.

GERIR A EMERGÊNCIA

Se o futuro for uma competição para alcançar uma inteligência única, perfeita e inquestionavelmente dominante, então parece provável que a humanidade perca o controlo de uma corrida existencial entre múltiplos atores ou sofra o exercício de uma hegemonia suprema por parte de um vencedor livre dos tradicionais travões e contrapesos. Quanto mais pequena for a margem pela qual a corrida é ganha, maior será a probabilidade de erro de cálculo humano.

A unipolaridade pode ser uma via para minimizar o risco de extinção. Se o atual líder fosse capaz de alargar a sua liderança até um limiar em que nenhuma outra entidade acreditasse ser possível reduzir a diferença, poderia haver uma maior certeza de garantir um certo nível de estabilidade. Pelo menos durante algum tempo, a base da ordem mundial – a procura incessante de equilíbrios transitórios e frágeis entre forças mais ou menos igualadas – poderia deixar de ser desejável.

Em alternativa, talvez fosse possível – embora dependente da supressão dos instintos humanos naturais – mediar um acordo entre entidades concorrentes para assegurar um determinado período de cooperação ou mesmo fundir num único empreendimento os múltiplos esforços dos concorrentes na cabeça da fila. Mas conseguir a unificação ultrapassando hostilidades geopolíticas e comerciais exigiria uma quantidade extraordinária de coragem e previsão. Qualquer ator que esteja na frente ou

perto dela espera ser incluído na companhia da vitória. Dentro dessa categoria privilegiada, os atores indecisos ou ainda atrasados seriam naturalmente os que mais beneficiariam com a participação no acordo, ficando bem posicionados para ganhar paridade e, eventualmente, avançar. Em contrapartida, outros atores principais poderiam considerar intolerável o sacrifício da sua posição de liderança, e a sua confiança no altruísmo poderia revelar-se demasiado frágil para resistir às tentações do abandono puro e simples.

Uma opção semelhante seria uma negociação para distribuir e delegar o poder. Quando confiante de que a vitória é iminente, o ator mais bem posicionado pode tentar convencer os rivais a render-se, garantindo-lhes acesso privilegiado a uma IA superinteligente. Mas nunca a confiança no empenhamento humano para com o seu semelhante, e muito menos para com o seu antigo adversário, durou para sempre. Se os humanos conseguissem, contra todas as probabilidades, forjar um tal acordo, está longe de ser claro como seria governada a sua implementação.

De facto, a simples recolha dos picos mais aguçados da «ilha» do conhecimento pode exacerbar a dinâmica que a sua consolidação tinha como objetivo evitar. A história da humanidade ainda não produziu qualquer registo de uma tal tentativa, muito menos de um êxito. Além disso, um tal acordo exigiria uma reorientação dramática da estratégia diplomática. Historicamente, a própria base da ordem mundial tem sido mantida pela procura incessante de um delicado equilíbrio entre forças paritárias. Neste caso, pelo contrário, os Estados-nações estariam a procurar uma estase hegemónica em grande parte alheia à prática humana. O bloco dominante de Estados seria provavelmente o criador original dos sistemas de IA e o fornecedor dos seus componentes mais decisivos, bem como o produtor do talento para os desenvolver e aperfeiçoar. A maioria não dominante poderia ser relegada para a condição de tributário, fornecendo dados e outros bens a troco de um certo nível de acesso às descobertas, sistemas de governação e defesas da IA.

Seja como for, não é este o cenário que preferimos ou previmos. Pelo contrário, acreditamos que não haverá apenas uma IA suprema e sim múltiplas instanciações de inteligência superior no mundo. Nesse caso, emerge um conjunto diferente de futuros possíveis. As nossas criações mais fortes, atuando como forças de compensação, poderiam estar mais bem

equipadas do que os humanos para exercer e manter um equilíbrio nos assuntos globais inspirado (mas não limitado) no precedente humano. A inteligência não humana poderia assim gerir a sua própria emergência, pelo menos nos domínios da segurança nacional e da geopolítica.

Mas poderá a IA criar uma divisão sustentável do domínio? Possivelmente, sim. As negociações diplomáticas, sob as camadas de emoção e psicologia humanas, são essencialmente uma espécie de teoria do jogo aplicada que é, ela própria, um ramo da matemática. Mesmo que a prática da diplomacia tenha nascido como uma forma de arte (ainda que exclusiva do domínio da conduta humana), pode tornar-se cada vez mais numa ciência. Ao fazê-lo, poderia transcender o historial bastante heterogéneo da diplomacia humana em matéria de identificar e perseguir oportunidades de compromisso. Os primeiros modelos de IA no Ocidente já se revelaram muito promissores na sua capacidade de aplicar a estratégia, pelo menos em termos de jogo, enquanto a China deu um passo em frente ao usar a inteligência das máquinas para desempenhar as funções dos diplomatas humanos.[64]

A tradição humana da diplomacia começou com a simples necessidade de transmitir mensagens entre sociedades de maneira segura e fiável. Gradualmente, os enviados de longe passaram a beneficiar de um tratamento especial, enquanto os Estados que violavam as novas normas raramente escapavam a sanções. Quando os representantes de Xerxes exigiram símbolos de submissão às cidades-Estado gregas, os diplomatas persas foram atirados para poços e fossos. Heródoto regista que mais tarde os deuses castigariam Esparta por este ato.[65] Cerca de dezassete séculos depois, quando um importante diplomata mongol foi executado por ordem de um xá persa, Gengiscão ordenou à sua temível horda que, como retaliação, destruísse todo o império corásmio.[66] Ainda assim, manter abertos os canais de comunicação, mesmo – e especialmente – em tempos de guerra, tornou-se um costume; com o tempo, criou-se um consenso básico de que talvez fosse melhor ouvir a mensagem do que matar o mensageiro.

Se as IA assumissem a condução da diplomacia entre si, poderiam ser intencionalmente treinadas para aceitar o mesmo costume, ou poderiam incidentalmente desenvolver uma tendência a favor da informação adicional que esse costume garantia. Continuaria certamente a haver aberrações.

Embora talvez não houvesse cabeças a rolar, as máquinas poderiam encontrar um modo equivalente de defender os seus estreitos interesses. Nesse caso, a sua racionalidade total, se os humanos conseguissem aceitá-la, poderia aumentar a probabilidade de, pelo menos, haver um começo mais seguro.

O fornecimento de soluções produzidas pelas máquinas para problemas humanos tão fundamentais como a diplomacia e a segurança levará naturalmente, no entanto, a uma maior dependência das capacidades da IA – uma dependência difícil de quebrar se e quando for necessária uma intervenção humana decisiva nos assuntos internacionais. Pelo menos a política humana, mesmo que imperfeitamente aplicada, impôs-nos a responsabilidade pelas nossas próprias escolhas. Optar por depender da IA, pelo contrário, pode reduzir a nossa capacidade de confiar no discernimento humano básico com o qual pudemos subsistir num passado mais previsível mas primitivo.

Porquê correr esse risco? Por um lado, para evitar o dilema da concorrência mutuamente destrutiva ou de uma hegemonia produzida por uma superinteligência emergente, mas também para nos protegermos de outras ameaças que se perfilam no horizonte. A cada ano que passa, a cada novo avanço tecnológico, diminui o limiar mínimo para a destruição de toda a sociedade humana.

A mitologia nórdica conta a história de Baldur, filho do deus Odin e da rainha Frigg, que as profecias da iminente e terrível morte do filho trazem alarmados.[67] A rainha Frigg, determinada a proteger o rebento de qualquer destino que se avizinhasse, viajou incansavelmente pelos nove reinos da Terra, lançando poderosas magias sobre todos os animais, elementos, plantas e pragas para que nenhum pudesse ser usado como arma contra ele. Mas o trapaceiro deus Loki, disfarçado de velha, conseguiu arrancar-lhe a informação de que o seu manto divino de proteção cobria tudo na Terra exceto o visco, a mais inofensiva de todas as ervas. Num banquete realizado para celebrar a imunidade de Baldur, os deuses, com a intenção de provar o bom trabalho da rainha, lançaram-lhe à vez armas de todos os géneros. Loki, no entanto, obrigou o seu irmão cego, Hodr, a disparar uma seta cuja ponta tinha sido esfregada com visco, perfurando o peito de Baldur e matando-o através do único objeto no Valhalla isento da em tudo mais abrangente tutela defensiva da rainha.

O significado deste mito antigo para a nossa situação moderna é simples e arrepiante: à medida que as ameaças se tornam cada vez mais disfarçadas e sofisticadas, as defesas da humanidade contra elas têm de ser cada vez mais perfeitas, dado que o mais pequeno erro ou omissão pode significar a derrota. E para atingir esse nível de perfeição, é muito possível que precisemos da ajuda da IA.

O exercício consiste, portanto, em decidir qual é o risco menor e, por conseguinte, qual deve preocupar-nos mais: sobreviver ao desenvolvimento da IA, sobreviver a outras revoluções paralelas como o advento da biologia sintética, ou sobreviver a potenciais catástrofes como as alterações climáticas radicais. É verdade que as IA mais pequenas podem ajudar na invenção imprudente de tecnologias existenciais; mesmo que os criadores destes novos meios fossem perfeita e adequadamente regulamentados, é muito provável que outros se revelassem menos cuidadosos e mais nocivos. Mas IA maiores poderiam ajudar na defesa contra essas mesmas tecnologias, permitindo uma verdadeira perfeição na tomada de decisões táticas defensivas.

Por exemplo, se desenvolvida, uma biodefesa individualizada – em que seriam introduzidos na nossa corrente sanguínea *nanobots* criados por IA com o objetivo de remover tudo o que não correspondesse a uma bioassinatura reconhecida – seria um corretor mais ágil do que a nossa atual resposta a ameaças biológicas. Do mesmo modo, a IA poderia gerar novos materiais e novos processos para reduzir as emissões de dióxido de carbono, reduzindo o risco de catástrofes climáticas.

É sem dúvida arriscado deixar que a IA assuma uma responsabilidade prematura e sustentada por espécies e sociedades para lá sua conceção, mas os caminhos tradicionais, que exigem perfeição no desempenho humano, podem ser ainda mais arriscados. O melhor, na nossa opinião atual, seria pôr as IA a trabalhar antes e não depois de a humanidade ter de confrontar a proliferação de novas ameaças à sobrevivência.[68] Assumindo isto, a pergunta pertinente é a seguinte: serão os humanos capazes de acelerar apenas as vias desejáveis da IA e protelar as indesejáveis?

A IA é uma força perturbadora não seletiva; a sua emergência, caso não seja gerida, representa um grande perigo para quem a cria e para quem a usa. É precisamente isto que pode obrigar rivais inicialmente relutantes a considerar acordos noutras circunstâncias implausíveis. Acreditamos que na

diplomacia, na defesa e talvez noutros domínios, alguns dos riscos da IA só podem ser geridos com êxito pela própria IA. A caixa de Pandora já foi aberta; mesmo que não tivesse sido, os benefícios da IA parecem ainda superar os riscos.

A nossa posição é, portanto, que o «mundo vulnerável» (para usar uma frase de Nick Bostrom) que criámos pode vir a exigir a intervenção da IA para sobreviver a alguns dos desafios e ameaças do desenvolvimento tecnológico da própria IA.[69] A questão mantém-se: como devem os humanos comportar-se perante um futuro que simultaneamente exige e proíbe o nosso controlo permanente?

UM NOVO PARADIGMA DE GUERRA

Durante quase toda a história da humanidade, a guerra foi travada num espaço definido, no qual se podia conhecer com razoável certeza a capacidade e a posição das forças hostis. A combinação destes dois atributos oferecia a ambos os lados uma sensação de segurança psicológica e de consenso comum, permitindo a contenção informada da letalidade. Só quando os líderes informados de cada umas das partes estivessem unidos na compreensão básica de *como* uma guerra podia ser travada, as forças em confronto podiam determinar *se* uma guerra devia ser travada.

Rapidez e mobilidade contam-se desde sempre entre os fatores mais previsíveis que sustentam a capacidade de qualquer peça de equipamento militar. Um dos primeiros exemplos é o desenvolvimento do canhão. Durante um milénio após a sua construção, as Muralhas de Teodósio protegeram a grande cidade de Constantinopla contra invasores externos. Então, em 1452, um engenheiro de artilharia do reino da Hungria – na altura sob suserania bizantina – propôs ao imperador Constantino XI a construção do basilisco, um canhão gigante que, disparado de trás das muralhas defensivas, pulverizaria os atacantes. Mas o despreocupado imperador, que não possuía os meios materiais nem a visão para reconhecer a importância da tecnologia, rejeitou a proposta.

Infelizmente para ele, o engenheiro húngaro revelou-se um mercenário político: mudando de tática (e de lado), atualizou o seu projeto para ser mais móvel – transportável por nada menos que 60 bois e 400 homens – e abordou o rival do imperador, o sultão turco otomano Mehmed II, que se preparava para sitiar a fortaleza inexpugnável. Conquistando o interesse do jovem sultão com a sua afirmação de que aquela arma seria capaz de «arrasar as muralhas da própria Babilónia», o empreendedor húngaro ajudou as forças turcas a derrubar os velhos muros em apenas 55 dias.[70]

Reencontramos frequentemente os contornos deste drama do século XV ao longo da história. No século XIX, a rapidez e a mobilidade transformaram a sorte em primeiro lugar da França, quando *la Grande Armée* de Napoleão subjugou a Europa, e depois da Prússia, sob a liderança de Helmuth von Moltke (o Velho) e de Albrecht von Roon, que tiraram partido dos recém-desenvolvidos caminhos de ferro e adotaram o controlo distribuído para permitir manobras mais rápidas e flexíveis. Do mesmo modo, a *blitzkrieg* – uma evolução dos mesmos princípios militares alemães – seria utilizada com grandes e terríveis efeitos contra os Aliados na Segunda Guerra Mundial.

A «guerra relâmpago» adquiriu um novo significado – e uma nova ubiquidade – na era digital. A velocidade é instantânea. Os atacantes não precisam de sacrificar a letalidade para manter a mobilidade, uma vez que a geografia já não é um constrangimento. Embora esta combinação tenha favorecido largamente a ofensiva nos ataques digitais, uma era da IA poderia assistir a um aumento da velocidade de resposta e permitir que as ciberdefesas evoluíssem a par das ciberofensivas.

Na guerra cinética, a IA provocará outro salto em frente. Os drones, por exemplo, serão extremamente rápidos e terão uma mobilidade inimaginável. Quando a IA for utilizada não só para guiar apenas um drone, mas frotas inteiras de drones, estas máquinas manobrarão em conjunto como um único coletivo coeso, perfeito na sua sincronia. Os enxames de drones do futuro dissolver-se-ão e reconstituir-se-ão sem esforço em unidades de todas as dimensões, tal como as forças de operações especiais de elite são constituídas por destacamentos capazes de ação autónoma com um comando próprio.

Por outro lado, a IA fornecerá defesas igualmente rápidas e flexíveis. Os enxames de drones são difíceis, se não mesmo impossíveis, de abater com projéteis convencionais. Mas armas dotadas de IA que disparassem fotões e eletrões (em vez de munições) poderiam recriar as mesmas capacidades destrutivas letais de uma tempestade solar, capaz de «fritar» os circuitos expostos de um satélite. Mais uma vez, a velocidade e a mobilidade situar-se-ão para lá da capacidade humana – e potencialmente igualadas na defesa e no ataque.

Com a velocidade e a mobilidade a deixarem de ser variáveis definidoras, a diferença de capacidade entre entidades concorrentes dependerá da precisão, do impacto imediato e da aplicação estratégica.

As armas dotadas de IA serão de uma exatidão sem precedentes. Há muito que os limites do conhecimento da geografia do inimigo restringem as capacidades e as intenções de todas as partes envolvidas num conflito. Mas a aliança entre a ciência e a guerra tem vindo a garantir uma precisão crescente dos nossos instrumentos, e é de esperar que a IA traga mais um ou muitos progressos nesse sentido. As IA irão reduzir a distância entre a intenção original e o resultado final, incluindo na aplicação de força letal. Seja através de enxames de drones baseados no solo ou a operar nos oceanos, ou possivelmente frotas interestelares, as máquinas serão capazes de matar seres humanos com uma precisão incrível, um escasso grau de incerteza e um impacto ilimitado. As fronteiras da potencial destruição dependerão apenas da vontade, e da contenção, dos humanos e das máquinas.

Assim sendo, a guerra na era da IA reduzir-se-á essencialmente a uma avaliação, não das capacidades do adversário mas das suas intenções e das aplicações estratégicas dessas intenções. Na era nuclear já entrámos de certo modo nessa fase, mas a sua dinâmica e o seu significado tornar-se-ão muito mais evidentes à medida que a IA provar o seu valor como arma. A questão fundamental é, pois, a seguinte: o que vão os comandantes equipados com IA querer e de que vão precisar?

Com uma tecnologia tão valiosa em jogo, não é provável que os seres humanos sejam considerados os principais alvos numa guerra de IA. Na realidade, as IA poderiam até eliminar por completo os humanos como seus procuradores no conflito, tornando-o menos mortífero mas potencialmente não menos decisivo. Do mesmo modo, parece improvável que o território,

por si só, provoque a agressão das IA – mas os centros de dados e outras infraestruturas digitais críticas poderão certamente fazê-lo. (É provável que os supercomputadores sejam escondidos e o armazenamento de inteligências distribuído, para garantir uma maior probabilidade de continuidade operacional e como defesa contra um «ataque de decapitação» mecânico).[71]

A rendição não vai, portanto, acontecer quando os efetivos de uma das partes estiverem reduzidos ao mínimo e o seu arsenal vazio, mas quando o escudo de silício dos sobreviventes se tornar incapaz de salvar os seus recursos tecnológicos e, finalmente, os seus delegados humanos. Os limites da destruição potencial dependerão apenas da vontade e da contenção do homem e da máquina. A guerra poderia evoluir para um jogo de baixas puramente mecânicas, sendo o fator decisivo a força psicológica do ser humano (ou da IA) que tem de competir para arriscar, ou desistir para evitar, uma instância de destruição total.

Até os motivos que regem o novo campo de batalha seriam em certa medida estranhos. G. K. Chesterton diz-nos: «O verdadeiro soldado luta não porque odeia o que está à sua frente, mas porque ama o que está atrás de si.»[72] É pouco provável que uma guerra de IA envolva amor ou ódio, e muito menos um conceito de bravura militar. Por outro lado, pode ainda incorporar ego, identidade e lealdade – embora a natureza dessas identidades e lealdades possa não ser consistente com a atual.

O cálculo na guerra sempre foi relativamente simples: o lado que primeiro considerar intoleráveis os danos causados pelo poder do antagonista fica sujeito à conquista e à conversão e só depois, talvez, à negociação. A consciência das próprias limitações produziu, de uma maneira fiável e natural, a contenção. Sem essa consciência, e sem a noção de dor (e portanto com total tolerância em relação a ela), não podemos deixar de nos perguntar o que, se alguma coisa, poderia levar à contenção uma IA que foi introduzida na guerra, e o que poderia pôr termo aos conflitos por ela travados. Iria uma IA que jogasse xadrez, se nunca tivesse sido informada sobre as regras que ditam o fim do jogo, continuar a jogar até ao último peão?

REESTRUTURAÇÃO GEOPOLÍTICA

Em todas as épocas da humanidade, quase como que em obediência a uma lei natural, surgiu, como um de nós disse certa vez, uma unidade «com o poder, a vontade e o ímpeto intelectual e moral para moldar todo o sistema internacional de acordo com os seus valores».[73] Na sequência do aparecimento dessa entidade, outras ligam-se em novos arranjos, criando dependências imprevisíveis em tempos de crise e ameaçando constantemente desfazer os equilíbrios geopolíticos de poder. Em alguns casos, o sistema resultante derruba as autoridades existentes; noutros, consolida-as.

O arranjo mais familiar das civilizações humanas é o sistema de Vestefália, tal como é convencionalmente entendido. A ideia do Estado-nação soberano tem, no entanto, apenas alguns séculos, tendo emergido de tratados assinados em meados do século XVII e que são no seu conjunto conhecidos como Paz de Vestefália. Não é a unidade preestabelecida de organização social e pode não ser adequada para a era da IA. De facto, à medida que a desinformação em massa e a discriminação automatizada desencadeiam uma perda de fé nesse acordo, a IA pode representar um desafio inerente ao poder dos governos nacionais. A desorientação psicológica e a possível retirada da realidade já descrita neste livro viriam provavelmente piorar ainda mais a situação. Em alternativa, a IA pode redefinir as posições relativas dos concorrentes no sistema atual. Se os seus poderes forem aproveitados principalmente pelos próprios Estados-nações, a humanidade poderá ser forçada no sentido de uma estase hegemónica, ou então de um novo equilíbrio de Estados-nações empoderados pela IA. Mas também pode ser o catalisador de uma transição ainda mais fundamental – uma mudança para um sistema totalmente novo em que os governos estatais seriam, por sua vez, forçados a abandonar o seu papel central na infraestrutura política global.

Uma possibilidade é as empresas que detêm e desenvolvem a IA virem a acumular um poder social, económico, militar e político totalizante. Os governos atuais são forçados a lidar com a sua difícil posição de anfitriões e líderes de claque de empresas privadas – usando o seu poder militar, capital diplomático e peso económico para promover os interesses domésticos

dessas empresas – e ao mesmo tempo o de apoiantes do cidadão comum, que desconfia da ganância e do secretismo monopolistas. Isto pode revelar-se uma contradição insustentável. E, como já foi descrito em pormenor, a emergência da IA tornará mais difícil a governação por qualquer instituição estabelecida.

Entretanto, as empresas poderão formar alianças para consolidar a sua já considerável força. Essas alianças podem ser construídas com base em vantagens complementares e no lucro da fusão ou, em alternativa, numa filosofia de desenvolvimento e implementação de sistemas de IA. Tais coligações empresariais poderão assumir as funções tradicionais do Estado-nação, com a diferença de, em vez de procurarem definir e expandir territórios delimitados, cultivarem redes digitais difusas como seus domínios.

E há ainda outra alternativa. A difusão descontrolada e de fonte aberta poderia dar origem a bandos ou tribos mais pequenas com uma capacidade de IA inferior mas substancial, suficiente para satisfazer as necessidades de administração, abastecimento e defesa dentro de um âmbito limitado. Entre os grupos humanos que rejeitam a autoridade estabelecida em favor de finanças, comunicação e governação descentralizadas, a proto-anarquia pode prevalecer. Ou esses grupos podem incorporar uma dimensão religiosa, talvez impulsionada por uma das conceções de IA e divindade exploradas num capítulo anterior. Afinal, em termos de alcance, o cristianismo, o islamismo e o hinduísmo têm sido maiores e mais duradouros do que qualquer Estado da história. Em épocas futuras, a denominação religiosa, e não a cidadania nacional, poderá revelar-se o enquadramento mais relevante para a identidade e a lealdade.

Em qualquer destes futuros, quer dominado por alianças empresariais ou disperso em agrupamentos religiosos livres, o novo «território» que cada grupo reivindicaria – e pelo qual lutaria – não seria centímetros quadrados de terra mas talvez engenhos digitais, como sinal de lealdade de utilizadores individuais. A ligação entre esses utilizadores e qualquer administração – inevitavelmente afetada pelo complexo efeito da IA no *status* do tradicional governo centralizado – iria subverter a noção tradicional de cidadania, e os acordos entre as entidades seriam diferentes das alianças habituais.

Historicamente, as alianças têm sido forjadas por líderes individuais e serviram para aumentar a força de uma nação em caso de guerra. Em contraste, a perspetiva de cidadanias e alianças – e talvez conquistas ou cruzadas – estruturadas à volta de opiniões, crenças e identidades subjetivas de pessoas comuns em tempo de paz exigiria uma nova (ou muito antiga) conceção de império. Também forçaria uma reavaliação das obrigações impostas por um juramento de fidelidade e o custo das opções de saída, se acaso as houver no nosso futuro emaranhado na IA.

PAZ E PODER

As políticas externas dos Estados-nações, e em consequência os sistemas internacionais, foram construídas e ajustadas equilibrando idealismo e realismo. Os equilíbrios temporários conseguidos pelos nossos líderes não são vistos, em retrospetiva, como resultados finais mas apenas como estratégias efémeras (ainda que necessárias) exigidas por uma dada altura. Com cada nova era, esta tensão produziu uma expressão diferente do que constitui ordem política. Um líder não pode limitar-se a concretizar uma opção que se enquadra num espectro existente e já considerado. Em vez disso, os governantes têm de fazer pelo menos algumas escolhas que decorram (ou pareçam decorrer) da inspiração – muitas vezes encorajando a prossecução de objetivos que se situam para lá do que na prática é possível alcançar.

A dicotomia entre a prossecução de interesses e a prossecução de valores – ou entre a vantagem de um determinado Estado-nação e o bem global – tem feito parte desta evolução sem fim. Na condução da sua diplomacia, os líderes dos Estados mais pequenos têm, historicamente, respondido de uma maneira direta e inequívoca, dando prioridade às necessidades da própria sobrevivência. Em contrapartida, os responsáveis por impérios globais, com os meios para atingir objetivos adicionais, enfrentaram uma situação mais angustiante.

Desde o início da civilização, a par com o crescimento, as unidades humanas de organização atingiram novos níveis de cooperação. Mas atualmente, talvez devido à escala dos nossos desafios planetários, bem como às desigualdades materiais evidentes entre os Estados e no seio deles, surgiu uma reação contra esta tendência. Poderão as IA mostrar-se à altura desta escala ainda maior da governação humana, capazes de ver com meticulosidade e fidelidade a interação do globo e não apenas os imperativos da nação? Será que se pode confiar nelas para calcular – mais precisamente do que alguma vez fizemos – primeiro os nossos interesses e os nossos valores, e depois a sua correta proporção e relação entre si?

Seria irrealista esperar, como um de nós já disse, que os líderes humanos «confinem as nossas ações a situações em que as nossas posições morais, legais e militares estejam completamente em harmonia e em que a legitimidade esteja mais de acordo com as exigências da sobrevivência».[74] Para os humanos, isto continua a ser verdade. Albergamos, no entanto, a esperança de que as IA, usadas para fins políticos no país e no estrangeiro, façam mais do que meros bons negócios. Idealmente, seriam capazes de oferecer-nos soluções novas e globalmente ótimas, atuando num horizonte temporal mais longo e com uma resolução mais elevada do que somos capazes, alinhando deste modo os interesses humanos diferentes e concorrenciais. No mundo que se avizinha, as inteligências artificiais que navegam no meio de conflitos e negoceiam a paz poderão ajudar a clarificar, ou mesmo a ultrapassar, os nossos dilemas tradicionais.

No entanto, se a IA viesse de facto a resolver problemas que devíamos ter resolvido sem ajuda, poderíamos enfrentar uma crise de confiança: isto é, excesso de confiança por parte de alguns e falta de confiança por parte de outros. Para os primeiros, uma vez compreendidos os limites da nossa capacidade de autocorreção, pode ser difícil admitir que acabámos por ceder demasiado poder à presumida sabedoria das máquinas para lidar com questões existenciais da conduta humana. Para os segundos, a verificação de que a simples remoção da ação humana da gestão dos nossos assuntos tem sido suficiente para resolver os problemas mais intratáveis que nos afligem pode revelar de maneira demasiado explícita as limitações da conceção humana. Se a paz sempre foi apenas uma simples escolha voluntária, o preço da imperfeição humana foi pago na moeda da guerra perpétua. Saber que uma solução sempre existiu mas nunca foi concebida

por nós seria esmagador para o orgulho humano. Este é um exemplo especialmente pungente do dilema da dependência – e subsequente perceção de inferioridade – explorado num capítulo anterior. Mas, no caso da segurança, ao contrário do que acontece com a nossa remoção do mundo científico ou empreendimentos académicos, podemos com mais facilidade reconhecer a necessidade de um mediador num divórcio litigioso. É nossa convicção, e nossa esperança, que, neste caso, alguns dos nossos piores traços de carácter nos permitam exibir alguns dos melhores: que o instinto humano para o interesse próprio, mesmo a expensas de terceiros, possa preparar-nos para aceitar a capacidade da IA de os transcender.

---

58 Paul Scharre, «America Can Win the AI Race», *Foreign Affairs*, 4 de abril de 2023.

59 William J. Broad *et al.*, «Israeli Test on Worm Called Crucial in Iran Nuclear Delay», *New York Times*, 15 de janeiro de 2011.

60 Entrevista com Dario Amodei, CEO da Anthropic por Dwarkesh Patel, https://www.dwarkeshpatel.com/p/dario-amodei#details.

61 Jeremy Hsu, «China's first underwater data centre is being installed», *New Scientist*, 4 de dezembro de 2023.

62 Walter Pincus, «Soviets Had Chance to Develop First A-Bomb, Historian Says», *Washington Post*, 27 de julho de 1979.

63 Graham Allison e Eric Schmidt, «Is China Beating the U.S. to AI Supremacy?», Projeto Evitar a Guerra Entre Superpotências no Harvard Kennedy School Belfer Center for Science and International Affairs, agosto de 2020, https://www.belfercenter.org/sites/default/files/2020-08/AISupremacy.pdf

64 A Google DeepMind e a Meta AI já criaram programas que dominaram os humanos no jogo da Diplomacia: Google DeepMind: János Kramár, Tom Eccles *et al.*, «Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy», *Nature*, 6 de dezembro de 2022; Meta: Meta Fundamental AI Research Diplomacy Team (FAIR), «Human-level play in the game of Diplomacy by combining language models with strategic reasoning», *Science*, 22 de novembro de 2022; a Academia Chinesa de Ciências foi mais longe, criando algoritmos de aprendizagem automática treinados em bases de dados governamentais que estão a ser utilizados pelos diplomatas chineses para a avaliação de riscos na verificação de projetos de investimento estrangeiro e na previsão de acontecimentos como convulsões políticas ou ataques terroristas; Stephen Chen, «Artificial intelligence, immune to fear or favour, is helping to make China's foreign policy», *South China Morning Post*, 30 de julho de 2018.

65 Ver Heródoto, *Histories of Herodotus*, trad. Henry Cary (Nova Iorque: D. Appleton and Company, 1904).

66 Ver Frank McLynn, *Genghis Khan: His Conquests, His Empire, His Legacy* (Filadélfia: Da Capo Press, 2015), 259.

67 Ver a parte conhecida como «Gylfaginning» (O Começo de Gylfi) em Snorri Sturluson, *The Prose Edda*, início do século XIII, apresentado como um diálogo entre a figura mitológica Gylfi, que representa um rei humano, e os três grandes deuses: Hárr, Jafnhárr, e **Þriði**.

68 Ver Flo Read, «Nick Bostrom: Will AI lead to tyranny?», *UnHerd*, 12 de novembro de 2023, https://unherd.com/2023/11/nick-bostrom-will-ai-lead-to-tyranny.

69 Nick Bostrom, «The Vulnerable World Hypothesis», *Global Policy*, vol. 10, n.º 4, novembro de 2019.

70 Roger Crowley, *Constantinople: The Last Great Siege 1453* (Londres: Faber and Faber, 2005), 91.

71 Ver o antigo conceito histórico e literário de um filactério, um «artefacto da alma» mágico que se diz ser usado por alguns feiticeiros para ancorar a alma ao mundo físico e reter a inteligência e a mente no caso de o corpo material ser destruído. O feiticeiro não pode ser morto definitivamente enquanto o seu filactério se mantiver intacto; por isso são normalmente escondidos. As histórias nórdicas (*Boots e os Seus Seis Irmãos*) falam de homens ou gigantes que escondem o coração noutro lugar, de modo a permanecerem imortais apesar de feridos em combate.

72 G. K. Chesterton, *The Illustrated London News*, 14 de janeiro de 1911, citado em https://www.chesterton.org/quotations/war-and-politics.

73 Henry A. Kissinger, *Diplomacy* (Nova Iorque: Simon & Schuster, 1994), 1.

74 Henry A. Kissinger, *Nuclear Weapons and Foreign Policy* (Nova Iorque: Harper & Brothers, 1957), 429.

CAPÍTULO 6

PROSPERIDADE

O ÉPICO NACIONAL finlandês *Kalevala* começa com Väinämöinen, o primeiro homem a trazer árvores e vida a um mundo até então estéril, a dar à praia na costa da distante Pohjola depois de ter sido derrotado numa terrível batalha naval.[75] Tendo, com os seus cuidados, restaurado a saúde do herói, Louhi, a perversa rainha que governa a negra e sombria terra de Pohjola, exige uma recompensa em troca da sua libertação. Não satisfeita com o ouro ou a prata, a velha bruxa do Norte pede o que então só existia no mito: o Sampo, uma máquina mágica capaz de proporcionar ao seu proprietário uma inesgotável fonte de riqueza.

Com o consentimento do irmão, Ilmarinen – arquiteto da cúpula do céu e única pessoa capaz de criar algo semelhante ao Sampo – Väinämöinen faz um acordo com a poderosa bruxa, prometendo que, se ela o libertar, honrará as suas dívidas enviando o irmão em seu lugar. Louhi seduz Ilmarinen prometendo dar-lhe uma das suas belas filhas virgens como noiva, e o mestre artesão faz com entusiasmo o que a rainha lhe ordena.

O ferreiro eterno convoca os ventos divinos para soprarem os foles durante três dias e forja com os melhores materiais do reino – «as pontas das penas de cisne-branco, o leite da maior virtude, um único grão de cevada e a mais fina lã de cordeiro» – o Sampo, enxertando-lhe nos lados torneiras das quais jorra um fornecimento infinito de cereais, sal e moedas.[76] Mas no momento em que Ilmarinen retira a máquina das chamas da forja, Louhi, vitoriosa, arranca-lha das mãos e tranca-a num cofre

escavado na montanha. Doravante, Pohjola prosperará graças à ilimitada capacidade de gerar riqueza do engenho, deixando Ilmarinen amargurado e deprimido.

Muitos anos mais tarde, Ilmarinen e Väinämöinen regressam juntos para corrigir a injustiça cometida pela bruxa. Quando chegam ao reino da abundância, ameaçam tomar o Sampo pela força se não lhes for dada metade dos lucros. Segue-se uma feroz batalha naval e, no meio do caos, o Sampo afunda-se nos negros abismos onde ainda hoje permanece, sem poder ser recuperado, a produzir riquezas sem dono e tornar salgada a água do oceano.

Há por todo o mundo histórias semelhantes a respeito de máquinas de abundância: o Akshaya Patra, um recipiente de cobre sem fundo descrito no épico hindu *Mahabharata*; o caldeirão mágico da abundância no mito irlandês do deus Dagda; o Uchide-no-kozuchi, martelo mágico do folclore japonês capaz de «bater» a pedido seja o que for, incluindo casas, roupas e até seres humanos.[77]

Aqueles que hoje constroem a IA acreditam que a sua criação vai ser esse celeiro perpetuamente abastecido, esse moinho mágico, essa cornucópia a transbordar de flores, frutos e trigo. Como a mitologia alerta, no entanto, a criação por si só não será suficiente. Para realizar o seu potencial, a IA vai ter, no seu desenvolvimento e implementação, de ser ajudada e associada a mudanças institucionais adequadas e a uma conceção política sensata. Deverá ser utilizada para afrouxar, e idealmente eliminar por completo, os laços de servidão que no passado dominaram as relações sociais e económicas da humanidade, e para nos lançar num futuro caracterizado por menos pobreza e desigualdade.

Este objetivo é extraordinariamente ambicioso seja por que medida for, e podemos, com razão, levantar questões. Mas e se a IA nos oferecer de facto uma ponte para uma nova idade de ouro? Mesmo um êxito parcial poderia constituir um renascimento civilizacional.[78]

CRESCIMENTO E INCLUSÃO

Em março de 2016, depois de perder três jogos consecutivos, Lee Sedol, o grande mestre coreano do jogo chinês Go, não sentiu raiva nem tristeza, mas espanto. Nunca imaginara que – tendo dedicado toda a sua vida ao domínio deste antigo jogo – pudesse ser derrotado pela IA, um adversário ultramoderno. E no entanto, apenas um jogo antes, o seu oponente computacional, conhecido como AlphaGo, tinha feito uma jogada – a número 37 – tão pouco ortodoxa que o obrigara a considerar uma vez mais que as máquinas podiam não só possuir uma capacidade bruta mas também poderes de criatividade.

Na altura, impressionado pelo desempenho do seu extraordinário competidor, Lee já não jogava apenas pela vitória – que já concedera na partida à melhor de cinco – mas por um final em beleza. No jogo seguinte, o quarto, respondeu à jogada 37 com a jogada 78: um golpe de génio que o levou, contra todas as probabilidades, à vitória naquele que continua a ser o único jogo que o AlphaGo alguma vez perdeu. Na Coreia e em todo o mundo, a comiseração deu temporariamente lugar aos festejos.

Durante aquela semana em Seul, o número um do mundo, apenas com uma pausa ocasional para fumar no pátio exterior do hotel de luxo que o patrocinava o encontro, travara uma luta solitária, envolvido numa competição com um adversário único que nunca esperara enfrentar jogando em nome de uma equipa – a humanidade – que nunca procurara representar. O confronto será sobretudo recordado, mais do que pelo resultado final, pela espantosa capacidade demonstrada pelos seres humanos: nomeadamente, o próprio Lee e os criadores da máquina que enfrentou.[79]

O lema da DeepMind, a empresa por trás da vitória na Coreia, é «Primeiro, resolver a inteligência; depois, usar a inteligência para resolver tudo o resto.»[80] A inteligência, como motor de nova criação, vai mudar a nossa compreensão de tudo. Perante este grande desconhecido, podemos sentir-nos derrotados. Mas, pelo menos em alguns contextos, poderá ser sensato reagir com a atitude demonstrada por Lee Sedol, que, ao enfrentar a IA, a tratou mais como uma inspiração do que como um rival.

A situação de Lee Sedol era única – a de especialista no topo da sua área a enfrentar IA no âmbito de uma experiência. Por isso pode estar naturalmente inclinado a assumir uma posição de admiração e imaginação em vez de ressentimento. Muitas outras pessoas reagirão à IA de maneira bem mais negativa.

Abster-se da rivalidade será particularmente difícil em situações que parecem ser de soma zero, talvez sobretudo na potencial deslocação do trabalho humano pela IA. Neste capítulo tentamos abordar essa presumida dinâmica de soma zero – que consideramos ser largamente mal interpretada – e descrever o que acreditamos poderem ser benefícios abundantes e amplamente distribuídos para a humanidade, mesmo, potencialmente, num mundo sem trabalho.

Para a maior parte dos seres humanos ao longo da maior parte da História o trabalho não tem sido um jogo que se espere ganhar nem uma forma de arte que se espere dominar, mas um fardo nada gratificante e brutal imposto por estruturas sociais que mantêm os trabalhadores presos ao serviço. Ainda que possam ter ajudado a manter a estabilidade, essas estruturas têm invariavelmente servido também para atormentar o espírito humano.

No *Bhagavad Gita*, que faz parte do *Mahabharata*, há um diálogo entre o príncipe guerreiro Arjuna e o seu cocheiro – a divindade Krishna disfarçada – que constitui um discurso pertinente sobre as hierarquias sociorreligiosas que durante muito tempo garantiram a ordem na sociedade indiana. Quando Arjuna hesita, no campo de batalha, em levantar a espada contra os seus parentes, Krishna explica-lhe em termos nada ambíguos que não pode haver desvio do dever, nem divergência do destino: «É melhor estragares o teu trabalho do que atingires a perfeição no do teu vizinho; morre se tiveres de morrer, mas não assumas o risco de trabalhos alheios.»[81]

Neste contexto, cada um desempenha um papel específico, por mais insatisfatório que seja, injustamente determinado pelo nascimento. Assim, de acordo com o *Gita*, «O dever dos brâmanes [a casta mais elevada] é ser pacífico e sábio; o dos soldados, lutar; o das classes médias cuidar da agricultura e do comércio, e o dos servos fazer trabalho braçal.»[82] Uma sociedade só pode ser bem-sucedida se todos cumprirem fielmente os deveres da sua condição. Os que o fizerem nesta vida têm uma hipótese de conseguir uma condição mais elevada na próxima; quanto aos que o não fizerem, espera-os o sofrimento na próxima reencarnação.

É claro que as castas hindus não estavam sozinhas. A teoria política de Aristóteles implicava papéis e deveres sociais rigorosos. A escravatura – apoiada na lei, na força e no tormento psicológico – tornou-se, em algumas

partes do mundo, a principal e cruel instituição para extrair trabalho e impor uma posição social fabricada.

Nos últimos dois séculos, as democracias capitalistas substituíram em grande medida a casta e o cativeiro por mercados de meritocracia; os eclesiásticos exaltaram e os académicos documentaram os valores sociais de uma forte ética de trabalho; e os trabalhadores adquiriram as artes da negociação e da greve. Mas mesmo assim, e quer os nossos trabalhos humanos sejam impostos pelo divino, pelo governo ou pela necessidade de garantir um salário, o nosso esforço mental e físico tem sido de um modo geral não tanto para nós mesmos como ao serviço de outrem.

Muitas guerras foram causadas por, ou resultaram de, alterações à pergunta «Quem recebe o quê – e porquê?» (para usar o título do *bestseller* de 2015 do economista Alvin Roth). A oferta relativamente fixa de terra, trabalho e capital garantiu que a escassez – e não a abundância – fosse o paradigma predominante da teoria e da prática económicas. Travam-se batalhas ferozes para decidir a maneira de dividir o que foi criado e, mais frequentemente ainda, sobre o modo de distribuir o pouco que resta. Estas fricções ocorrem tanto no seio das sociedades como entre elas, mesmo em tempo de paz, com os cidadãos a debaterem as raízes da desvantagem relativa e a exigirem a redistribuição para resolver o facto do sofrimento generalizado.

Um aumento do montante total de riqueza disponível para redistribuição e, subsequentemente, um aumento efetivo do volume de riqueza redistribuída, elevaria o nível de vida das pessoas em todo o mundo. Se realizado a uma escala enorme – a escala necessária para convencer qualquer sociedade, ou qualquer entidade dentro de uma sociedade, da suficiência da sua riqueza – um tal desenvolvimento poderia transcender as discussões contemporâneas sobre o sustento e centrar a nossa atenção na abundância.

A IA representa uma oportunidade real de deslocar pelo menos um dos fatores de produção originais, transferindo a função do trabalho dos seres humanos para as máquinas. Além disso, será utilizada para investigar e desenvolver fontes de matérias-primas cada vez mais baratas e abundantes para os seus próprios fatores de produção. À medida que for sendo simultaneamente utilizada no fabrico, poderá reduzir o capital necessário para produzir um dado bem. É verdade que continuarão a ser necessários

alguns elementos e matérias-primas não renováveis para equipar a própria inteligência não humana, mas isso pode mudar se a IA for usada com êxito para encontrar ou gerar substitutos sintéticos. Uma nova arquitetura de computação, mais eficiente do que a atual por várias ordens de grandeza, poderia ser redesenhada pela IA e, eventualmente, também as fábricas que produzem os componentes constituintes da própria IA.

Ao criar substitutos sintéticos mais sustentáveis para uma grande variedade de bens, a IA poderia permitir uma nova era de abundância. Mesmo tendo em conta alguns constrangimentos físicos e materiais atuais, os seus contributos – embora dificilmente infinitos – poderiam ser de tal magnitude que satisfizessem todas as necessidades básicas da humanidade e concretizassem muitas das suas esperanças. Isto pode aliviar a pressão que o paradigma da escassez tem exercido sobre a nossa psicologia, bem como o pessimismo induzido pela obrigação de trabalhar como meio de sobrevivência.

Sam Altman, diretor-executivo da OpenAI, analisou os sistemas económicos segundo duas variáveis: crescimento e inclusão.[83] Muitas sociedades conseguiram alcançar uma ou outra, pelo menos durante um certo período; muitas menos conseguiram manter ambas de modo consistente. Sobre esta relação, Altman escreve:

> O capitalismo é um poderoso motor de crescimento económico porque recompensa as pessoas por investirem em ativos que geram valor ao longo do tempo, o que constitui um sistema de incentivos eficaz para criar e distribuir ganhos tecnológicos. Mas o preço do progresso no capitalismo é a desigualdade.

Por outras palavras, a IA e os ganhos de produtividade que lhe estão associados podem catalisar naturalmente um período prolongado de crescimento. Mas a inclusão só ocorrerá por opção.

Por conseguinte, num mundo pós-IA, talvez a solução fosse, como sugere Altman, tributar os dois «ativos que constituirão a maior parte do valor [no] mundo», nomeadamente as empresas – sobretudo as que constroem, mantêm e utilizam a IA – e a terra, que ainda permanece fixa (pelo menos aqui na Terra). Claro que se os seres humanos não são individualmente responsáveis pelo valor do trabalho que resulta dos

conhecimentos gerados pela IA, é lógico que esse valor seja partilhado. E é muito possível que a terra (e, durante algum tempo, os minerais de terras raras essenciais para a computação) se conte entre os poucos ativos verdadeiramente fixos – e, portanto, valiosos e, portanto, tributáveis – num mundo pós-escassez.

Esta sugestão pressupõe, no entanto, a continuação da existência de nações (ou de substitutos reconhecíveis) como agentes de redistribuição, e de empresas como potenciais alvos de tributação. Além disso, atribuir uma mais-valia tanto à terra como à inovação poderia incitar a lutas intermináveis e possivelmente violentas pela primazia. Uma visão alternativa para garantir a equidade na era da IA poderia ser encontrada numa função análoga à do mercado de ações: nomeadamente, a criação e atribuição global automática de unidades divisíveis de riqueza associadas aos lucros crescentes dos modelos de IA. (Os direitos monetários poderiam vir acompanhados de direitos de voto, como acontece com algumas ações).

Ainda outra possibilidade seria concentrarmo-nos menos na propriedade da IA e mais na distribuição dos seus benefícios finais. Mas isto não poderia ser feito sem oposição, tanto pelos motivos óbvios de que se devia, em vez disso, distribuir a propriedade dos meios de produção, como pelos motivos práticos de que seria necessária uma enorme quantidade de logística e monitorização para garantir um determinado padrão em milhares de milhões de indivíduos.

Outra opção poderia inspirar-se no sistema de patentes: permitir a propriedade exclusiva de uma invenção de IA e dos seus lucros, de modo a incentivar a melhoria, mas apenas por um período limitado; após uma determinada data, assumindo que a sua segurança foi comprovada, o modelo poderia ser publicado (ou, talvez, feitas cópias e construídas infraestruturas em novos locais) para utilização comum, replicação e ganho.

Em resumo: os pioneiros da IA podem subestimar o alcance dos desafios económicos e políticos que puseram em marcha. Sim, a IA será capaz de fazer praticamente tudo. Mas, como Sam Altman diz sem rodeios: «Vai fazer o que eu quero, ou vai fazer o que tu queres?»[84] Como é que «nós» vamos decidir, e quem é «nós»? Gerir estas energias de enorme possibilidade e redistribuir os benefícios dessa gestão é uma grande responsabilidade. Os futuros decisores devem ter o cuidado de não enraizar

mais uma vez o tipo de desigualdades sociais e económicas que se propagaram a partir da Revolução Industrial antes de começarem a ser corrigidas, demasiado lentamente, através de estruturas de controlo mais dirigidas pelo homem.[85] [86]

Atualmente, a preponderância dos benefícios das IA avançadas e o controlo quase exclusivo sobre elas têm sido atribuídos a um número extremamente reduzido de indivíduos. Irão eles desistir das suas vantagens? Se o fizerem – se e quando mais benefícios e mais controlo passarem a ser partilhados ao nível nacional – começarão imediatamente a surgir os apelos à globalização de ambos. Irá uma nação transformar a sua riqueza soberana em benefício comum? Alguns argumentarão que as barreiras psicológicas ao sacrifício desaparecerão quando o mundo deixar de ser de soma zero. Mas isto pressupõe uma transição que ainda não aconteceu, uma transição que parece contrária ao atual *status quo* e que – se vier a acontecer – vai ter de ser o resultado de uma escolha humana.

Além disso, mesmo que consigamos chegar a um mundo para lá da escassez, não é claro como seriam estruturados os incentivos humanos globais de modo a permanecerem em harmonia. Um mundo para lá do valor não é um mundo para lá dos valores. Os que não comungam com base no dinheiro podem comungar – e depois proceder à reorganização da sociedade e das instituições globais – com base na religião, na raça, na linhagem familiar, na educação, na moralidade, na habilidade, na estética, no humor ou em qualquer outra coisa. Os que não lutam por dinheiro podem lutar por Deus, pelo poder, pela glória ou pela vingança.

Mais, a história económica confirma a dificuldade de conceber sistemas que sejam tão coerentes quanto são eficazes. E é preciso admitir que os humanos têm sido ineptos na previsão dos efeitos a longo prazo da tecnologia; no caso da IA, o nosso otimismo pode ser injustificado e as nossas apreensões descabidas.

Seja como for, os autores deste livro acreditam que a IA pode ser aproveitada para gerar uma nova base de riqueza e bem-estar humanos – e essa possibilidade por si só exige que comecemos a avançar nesse sentido. Além disso, estamos confiantes de que, se esse esquema económico e político se concretizar, irá pelo menos aliviar, se não eliminar, as tensões de trabalho, classe e conflito que anteriormente dilaceraram a humanidade.

MOBILIDADE

Se, na era da IA, a nossa taça transbordar, como garantir que todos os seres humanos beneficiam à distância e de modo fiável do novo excedente? A humanidade já fez grandes progressos na distribuição de valor: desde as primeiras moedas, que aliviaram os encargos anteriormente impostos pelos sistemas de troca, até às replicações posteriores de moedas fiduciárias em papel e metal, bem como em invenções digitais como cartões de crédito, transferências eletrónicas e serviços bancários móveis. Apesar de terem facilitado um movimento mais eficiente de valor através do espaço, tiveram, devido à instabilidade da oferta, um desempenho menos bom quanto a capacidade ao longo do tempo. Os modernos projetos de moeda procuram a transferência eficiente de valor através do espaço e do tempo.

A economia moderna continua a ser formada pela produção de bens e pela prestação de serviços, e não pelas suas representações indiretas e abstratas. O dinheiro é nulo sem um mercado; torna-se apenas uma entrada numa base de dados para a afetação de recursos, na qual o valor deixa de ter significado. Visto através da lente da teoria da informação, o dinheiro tem hoje muito em comum com uma ligação à Internet, que para qualquer uso reconhecível exige também um contexto relacional. Como otimizar da melhor maneira os seus atributos técnicos pode vir a ser, na era da IA, uma tarefa de enorme urgência filosófica e tecnológica. Muito possivelmente, o mundo vai precisar de um novo género de rede financeira que equilibre as funções tradicionais do dinheiro como um consistente repositório de valor e meios de troca convenientes.

A IA pode também trazer uma nova dinâmica aos mercados financeiros e às políticas económicas. É fácil imaginar, abstratamente e em princípio, o que a IA pode fazer para criar riqueza – mesmo com acesso apenas através da Internet. Mas ainda que inventássemos moedas, sistemas, mercados e políticas que respondessem de uma maneira sensata à emergência de uma tal escala de potencial criação de valor, como iria a IA exatamente resolver ou eliminar, por exemplo, o problema da pobreza? Em termos práticos, como estabeleceria uma linha de base absoluta e global para a nossa qualidade de vida?

Se a IA se mostrar capaz de distribuir fisicamente os bens de que os humanos precisam para as suas necessidades materiais fundamentais, o volume de materiais que pode ser produzido e transportado à volta do mundo será algo sem precedentes. Há indiscutivelmente muitas possibilidades para um tal empreendimento – e outras tantas armadilhas. Podem ser concebidos sistemas de IA distribuídos que se conectam, através de robôs produzidos em massa e de sistemas de infraestruturas otimizados por IA, aos que não estão conectados. Essas novas ligações – que atualmente faltam a cerca de 2,6 mil milhões de pessoas – permitiriam fornecer alimentos, vestuário e abrigo à metade da humanidade que ainda não tem satisfeitas essas necessidades básicas.[87] Em alternativa, utilizando novos materiais sintéticos sustentáveis, as IA poderiam construir cidades em todo o mundo para proporcionar abrigo, regular a temperatura, garantir o acesso à energia e a ligações digitais e fornecer água potável, alimentos, medicamentos e saneamento. As inteligências das máquinas com graus diversos de complexidade poderiam inclusivamente ser prestadoras de serviços nessas cidades com capacidade para albergar dezenas de milhões de pessoas que, de outro modo, viveriam fora do sistema de prosperidade de que muito poucos de nós desfrutam nos dias de hoje.

Talvez ambas as alternativas se concretizem, juntamente com outros futuros que ainda não imaginámos. Nesse caso, os seres humanos manterão as suas opções – e provavelmente gozarão de maior liberdade de escolha do que antes. Com os seres humanos já não restringidos pela geografia do nascimento, pelas comunidades dos parentes ou pelos mercados de trabalho para as suas competências, para onde se deslocariam se a imigração deixasse de ser uma trágica necessidade para alguns e passasse a ser uma escolha livre para todos?

Se a IA construísse um grande número de novas cidades com a mesma elevada qualidade, continuaríamos a assistir a uma migração em massa das nações em desenvolvimento para as desenvolvidas, ou dos meios rurais para os urbanos, como tem acontecido historicamente?

Uma maior liberdade psicológica na constituição de famílias poderá também tornar-se uma realidade graças à IA. Alguns poderão optar por ter mais filhos, libertados da necessidade de tomar decisões difíceis sobre a afetação de recursos com base nas melhores hipóteses de sobrevivência e êxito. Outros, já não dependentes da descendência para garantir o sustento

na velhice, poderão abster-se de procriar. Os encargos geracionais poderão dissipar-se, permitindo que os filhos procurem áreas geográficas e profissões que anteriormente talvez tivessem descartado.

Os anteriores desequilíbrios e assimetrias na economia global resultaram de diferenças, entre outros fatores, da posse de recursos, da geografia e do capital humano. A IA poderia reduzir as lacunas de talento e igualar a distribuição de recursos. Isso tornaria menos relevantes as linhas de fratura predominantes ao longo das quais temos conduzido as trocas e o comércio global e que definiram a estratificação da prosperidade no mundo. Os países que hoje estão em desvantagem devido à sua pequenez territorial ou a qualquer outro motivo, ou que sofrem os efeitos da «fuga de cérebros», poderão encontrar novos meios para atingir os mesmos padrões circunstanciais de que gozam os gigantes económicos tradicionais do globo.

Como passar da atual desigualdade internacional para o futuro que descrevemos? O primeiro passo será talvez a conceção de sistemas e aplicações capacitados pela IA para fazer avançar as ciências materiais e otimizar as ligações digitais – incluindo os conjuntos de dados especificamente criados para o efeito que permitiriam a esses sistemas funcionar numa variedade de contextos globais. Ao investirmos nestas ideias, devemos ter em mente os ganhos potenciais: uma transformação profunda no padrão geral da vida humana coletiva e na igualdade das vidas individuais no que se refere a raça, género, nacionalidade, local de nascimento e antecedentes familiares. Uma igualização do custo da inteligência, distribuída por todo o globo, poderia criar um campo de jogo nivelado como nunca existiu.

ABUNDÂNCIA SEM ABANDONO

Mas e se a IA, ao mesmo tempo que funciona como um igualizador económico, fizer o custo da inteligência e, portanto, do trabalho, cair para zero? Isso encerraria o breve, mas maravilhosamente produtivo, período da história humana que permitiu aos indivíduos das sociedades livres melhorar

as suas condições de vida, caso fosse essa a sua escolha, através do próprio esforço. Tal como a escassez tem sido o paradigma do passado, a competição – pelo menos na era moderna – tem sido a condição por defeito da auto-organização, levando naturalmente a grandes variações na distribuição dos resultados em função da ambição, da capacidade, e da lotaria do lugar onde se nasceu e de quem se é filho.

Tudo isto significa que, de um modo geral, alguns seres humanos mais diligentes na execução do seu trabalho ou no aproveitamento do trabalho alheio se têm saído melhor, e outros pior. Se, no entanto, eliminarmos a função de seleção do trabalho, teremos também de enfrentar a eliminação das profissões e do *status*, identidade e significado que lhes estão associados. Seria de facto um mundo diferente.

Os nossos instintos naturais para superar a adversidade, para celebrar a excelência, para nos orgulharmos da diferença e da diversidade assinaláveis, continuarão certamente a existir, ainda que obrigados a procurar novos canais. Tal como outrora houve distribuições desiguais de talentos humanos para o trabalho, podem surgir novas distribuições desiguais de talentos humanos para o lazer. E, desta vez, essas distribuições podem não encaixar nos eixos de capacidade existentes, mas em eixos de uma qualidade diferente: curiosidade, sobriedade, bondade, ou talvez outra coisa completamente nova.

Num mundo sem trabalho, muitos de nós podem deixar-se seduzir e prender por mundos imersivos simulados e personalizados: uma sinfonia dos sentidos de visão, som, cheiro, tato e até paladar, já possível graças aos poderes completos da IA no domínio virtual. Como dissemos no Capítulo 3, milhares de milhões de humanos passivos poderiam plausivelmente escolher – ou ser atraídos por – este caminho, não lhes parecendo fácil nem necessário resistir a um *upgrade* tão instantâneo do nível de estimulação da sua realidade e da sensação de controlo sobre ela.

Mesmo hoje, quando as pessoas não estão ocupadas com as tribulações da vida e do trabalho, é provável que a luz em muitos olhos reflita o brilho de placas de vidro pixelizadas. Em todos os cantos do mundo, os trabalhadores, esgotados pelo trabalho árduo, retiram-se compreensivelmente para este meio mais leve e puro para consumir e criar.[88] Se os seres humanos já lutam para conter a sua atração pelas tecnologias relativamente primitivas dos nossos dias, como iremos lidar

com as imensamente superiores «máquinas de experiências» que a IA virá certamente possibilitar?[89] Se muitos de nós mal conseguem tolerar o padecimento do trabalho, como resistiremos ao prazer irrestrito?

A resposta é que a psicologia humana vai ter de evoluir em paralelo com a IA e os seus efeitos. É difícil prever exatamente como, mas parece verosímil que a IA possa de facto promover o significado humano tanto ou mais do que o diminui. O prazer por si só não satisfará o nosso desejo inato de significado. O trabalho – mesmo que não remunerado – pode proporcionar um sentido de propósito, desde que os meios exijam esforço ou os fins sejam nobres.

Uma experiência pessoal de dificuldade pode incutir um sentimento de orgulho quando o desafio é ultrapassado. O trabalho árduo, especialmente quando combinado com o empenho, empresta o seu próprio arco narrativo à nossa compreensão do tempo, do eu e da capacidade de domínio humana. Tendo em conta a nossa psicologia, muitas fontes de alegria e contentamento podem perfeitamente permanecer inalteradas na era da IA, apesar da miríade de mudanças que vão de certeza revolucionar as nossas vidas.

Não quer isto dizer que vamos simplesmente voltar às atividades do passado; em vez disso, podemos descobrir aspetos do potencial humano que antes não tínhamos verdadeiramente procurado. Consideremos as possibilidades que podem emergir de períodos de concentração humana até agora impossíveis de incluir na semana de trabalho de uma pessoa comum. Exercícios mentais e espirituais, praticados de uma maneira consistente, podem elevar a consciência humana. Períodos prolongados de consciência elevada podem, por sua vez, ajudar as nossas ligações relacionais com outros seres humanos (e animais), reforçar as nossas perceções do divino e produzir níveis significativamente elevados de bem-estar individual.

Os seres humanos que, sem ajuda, realizam proezas aparentemente sobre-humanas, sobretudo as que envolvem o uso do corpo físico, continuarão sem dúvida a ser casos fascinantes. Com um maior número de seres humanos a participar e a tentar dominar essas atividades, é de esperar que o nível de excelência aumente. Os desportos e jogos que levam o ser humano ao limite poderão crescer em prevalência e qualidade. A arte poderá florescer, porque o toque do autêntico conservará provavelmente o seu encanto.

Historicamente, as universidades têm feito questão de introduzir de modo igual os seus estudantes nas ciências e nas humanidades.[90] Acreditamos que estes empenhamentos quintessencialmente humanos, cada um a seu modo uma «demanda de significado» (para usar uma frase de um líder espiritual)[91], só poderão expandir-se. No Ocidente, as artes liberais eram as disciplinas e as competências consideradas pela antiguidade clássica como apanágios de uma mente independente. Na antiga Sinosfera, o artista Zhang Yanyuan, da dinastia Tang, estabeleceu que um «cavalheiro erudito» devia ser proficiente nas «quatro artes»: capacidades de som, visão, estratégia e escrita.[92] No futuro, e em contraste com as escolas de educação vocacionais, profissionais ou técnicas, podemos esperar um renascimento de algumas das primeiras tentativas de criar o «indivíduo erudito» capaz e desejoso de vocações diversas. As escolas de todo o mundo poderão produzir os filósofos e escritores de que precisamos para nos reorientarmos numa era totalmente nova.

Disciplinas outrora reservadas a uns poucos privilegiados poderão tornar-se o padrão para muitos, substituindo o anterior foco da sala de aula média na montagem do trabalho produtivo. A multiplicação de educadores de IA poderá possibilitar o ensino individualizado e seminários socráticos em todo o mundo. Imaginemos que, tal como o jovem Albert Einstein foi discípulo de Max Talmud (mais tarde Max Talmey), Voltaire do Abbé de Châteauneuf e Ada Lovelace (que escreveu o primeiro algoritmo de computador) de Mary Somerville, todas as crianças passariam a estar igualmente capacitadas para dominar a mente e o carácter.[93]

Podemos imaginar que, no lugar onde hoje se encontram as universidades – fiéis à sua forma original, um conjunto de dormitórios que rodeiam uma biblioteca em torno da qual se reúnem mentes ávidas de alcançar e fazer avançar a fronteira intelectual – os campus poderão um dia incluir espaços de congregação para equipas de humanos dedicadas a interpretar as descobertas da própria IA: isto é, compreendê-las e traduzir as mais destacadas em relevância para a vida humana.

Neste novo ramo das ciências, os seres humanos poderão optar por coevoluir com a IA, para continuarem a ser parceiros das máquinas na fronteira. Ou isso pode até acabar por não ser necessário para maximizar os benefícios das descobertas da IA para a humanidade. Em qualquer dos casos, e especialmente no último, esperamos que esta seja uma atividade

extraordinariamente difícil, trabalhando lado a lado com as nossas máquinas em turnos ininterruptos. Mas, por muito cansativo que seja, o esforço será essencial.

O PRIVILÉGIO DA ESCOLHA

Perante a ameaça percebida da automatização do trabalho humano, muitos comentadores fixam-se hoje no advento de uma nova crise espiritual: num mundo de abundância partilhada, argumentam, tornar-nos-íamos como os irresponsáveis vencedores de lotarias, dominados pelo hedonismo do excesso. Esta é, a nosso ver, uma perspetiva de privilégio. Para apreciar o extraordinário bem que a IA fará a milhares de milhões de pessoas – incluindo aquelas que atualmente não têm o dinheiro, nem as ligações, nem a necessidade básica e o tempo de lazer para participar em conversas como esta – basta recordar o destino contrastante dos seus antepassados: gerações de pais a trabalhar com máquinas de fabrico primitivo, mães a labutar em campos de escassas colheitas, crianças-operárias às quais foi roubada a inocência. Se, amanhã de manhã, todos os seres humanos pudessem optar por deixar de trabalhar, suspeitamos que a maioria o faria, enquanto os poucos que recusassem teriam provavelmente o privilégio de trabalhar não por obrigação mas por opção. A IA poderia ser direcionada para fazer o que já não *temos* de fazer, precisamente para podermos fazer o que *queremos* fazer.

Tal como se explica no Capítulo 3, receamos que uma grande fração dos seres humanos se possa tornar principalmente consumidora passiva de conteúdos gerados pela IA. Mas esta é uma preocupação que deriva da tendência humana para o consumo fácil – e, mais preocupante ainda, daquilo que só podemos supor que será a perceção futura que a IA terá de nós. Por outras palavras, a nossa preocupação com a passividade humana não tem que ver com a perda do trabalho remunerado. Já temos um protótipo de como as pessoas vivem quando podem ter o que querem sem trabalhar. Chamamos-lhes os ricos e os reformados. É certo que os ricos – incluindo muitos que não eram ricos anteriormente – por vezes não sabem

como exercer as suas opções depois de tantos anos a acumulá-las. Como Tolstoi confessou: «Se uma fada me aparecesse e se oferecesse para realizar todos os meus desejos, não saberia o que desejar.»[94]

A adaptação à abundância é provavelmente um problema de transição e não um desafio permanente. Alguns sentirão inicialmente que a introdução do trabalho automatizado os priva da sua principal fonte de realização e alegria. O que será sem dúvida uma experiência chocante. Mas parece-nos provável – não como resposta à nossa exortação, mas como resultado do instinto humano – que, se lhes for dado tempo, os humanos escolherão perseverar, talvez em novas vias ou como parceiros da IA, evitando a atrofia e destacando-se em vez disso como pensadores e fazedores. Em última análise, se estabelecermos os sistemas necessários para a distribuição, conexão, participação e educação, os seres humanos – empoderados e inspirados pela IA – podem continuar a trabalhar não por dinheiro, mas por prazer e orgulho.

---

75 Elias Lönnrot, *The Kalevala*, 1835. Compilado a partir de antigas baladas fino-carelianas, canções líricas e encantações que faziam parte da tradição oral finlandesa.

76 Hanna-Ilona Härmävaara, «The myth of the Sampo – an inifite source of fortune and greed – Hanna-Ilona Härmävaara», TED-Ed animation, 23 de setembro de 2019, https://www.youtube.com/watch?v=71fLFOjruFc.

77 *Mahabharata*, «Adi Parva» («Livro do Início»), «Vana Parva» ou o «Livro da Floresta». Composto por volta do século IV a.C. O Caldeirão do Dagda aparece no conto épico irlandês «A Segunda Batalha de Mag Tuired», um texto medieval irlandês provavelmente composto por volta do século XI ou XII. O «martelo mágico» tem origem no conto popular«Uchide-no-Kozuchi», que se traduz por «O Pequeno Martelo Mágico», associado ao lendário herói Urashima Taro. Registada e compilada durante o período Edo (1603-1868) ou antes.

78 Ver Sam Altman, «Moore's Law for Everything», 16 de março de 2021, https://moores.samaltman.com.

79 Ver o filme-documentário *AlphaGo – The Movie*, produzido por Greg Kohs, 13 de março de 2020, https://www.youtube.com/watch?v=WXuK6gekU1Y.

80 Tom Simonite, «How Google Plans to Solve Artificial Intelligence», *MIT Technology Review*, 31 de março de 2016.

81 Arthur W. Ryder, *The Bhagavad-Gita* (Chicago: University of Chicago, 1929), 3:15.

82 *Ibid.*, 18:41-44. Ver James Hijiya, *The Gita* de J. Robert Oppenheimer, *Proceedings of the American Philosophical Society*, vol. 144, n.º 2, junho de 2000.

83 Sam Altman, «Moore's Law for Everything.»

84 Ross Andersen, «Does Sam Altman Know What He's Creating?»

85 Ver Daron Acemoglu, *Power and Progress* (Nova Iorque: PublicAffairs, 2023) para uma discussão sobre o modo como a riqueza gerada pelas tecnologias-chave da Revolução Industrial reverteu inicialmente a favor de apenas alguns países e indivíduos. Acemoglu argumenta que os seres humanos foram, na realidade, bastante criativos e inventivos durante o período medieval com, por exemplo, muitas inovações na agricultura e no comércio, mas a visão prevalecente desse período era a de uma pequena elite que se afirmava detentora de um poder divino, açambarcava todos os proventos de uma maior produtividade e os aplicava na construção de grandes catedrais (que não aumentavam significativamente a produtividade, a saúde pública, etc.).

86 Tem sido sugerido, por exemplo, que certas estruturas governativas empresarializadas sejam obrigadas a ter em conta considerações não monetárias.

87 International Telecommunication Union (ITU), «Population of global offline continues steady decline to 2.6 billion people in 2023» (Comunicado de imprensa: Genebra, 12 de setembro de 2023). https://www.itu.int/en/mediacentre/Pages/PR-2023-09-12-universal-and-meaningful-connectivity-by2030.aspx.

88 Jay Olson *et al.*, «Smartphone addiction is increasing across the world: A meta-analysis of 24 countries», *Computers in Human Behavior*, 129 (2022), 107138.

89 A designação «Máquinas de Experiências» foi tirada de Robert Nozick, *Anarchy, State, and Utopia* (Oxford, UK: Blackwell, 1974), 42.

90 Inspirado na carta de fundação da Universidade de Stanford (https://www.stanford.edu/about/history/): «As universidades são uma multiplicidade de institutos, escolas, laboratórios e departamentos que se fertilizam reciprocamente com ideias e inovações. Exploramos – na escola de ciências – o código molecular que nos torna humanos, e – nas escolas de letras – a cultura que é igualmente essencial para a humanidade. Sustentado pelo empenhamento da universidade na busca e valorização do conhecimento...»

91 Viktor Frankl, *Man's Search for Meaning* (Boston: Beacon Press, 2006), 6: «Life is not primarily a quest for pleasure, as Freud believed, or a quest for power, as Alfred Adler taught, but a quest for meaning.» Encontrado num prefácio de Rabbi Harold Kushner.

92 As «quatro artes» (*si yi*) são *qin* (um instrumento de cordas), *qi* (o jogo de estratégia Go), *shu* (caligrafia), and *hua* (pintura chinesa).

93 Ver Erik Hoel, «Why we stopped making Einsteins», *The Intrinsic Perspetive Substack*, 16 de março de 2022, https://www.theintrinsicperspective.com/p/why-we-stopped-making-einsteins.

94 Lev Nikolayevich Tolstoi, *A Confession and Other Religious Writings*, trad. David Patterson (Nova Iorque: W. W. Norton, 1983), 28.

CAPÍTULO 7

CIÊNCIA

A IA está pronta para redefinir as possibilidades de qualquer empreendimento criativo e procurar novas conclusões em todos os domínios científicos. Por sua vez, é quase certo que as explorações subsequentes refinarão e alargarão o círculo da compreensão humana. Tal como dissemos no Capítulo 1, é concebível, talvez até provável, que a IA se projete em todas as direções ao mesmo tempo e que os seus êxitos em todas as frentes sejam validados, absorvidos e combinados em grandes blocos de novos conhecimentos humanos.

Se o passado recente foi definido pelos triunfos do homem na engenharia de sistemas complicados – microprocessadores, a Internet, motores a jato, aceleradores de partículas – o futuro será definido pelos progressos da IA na engenharia de sistemas complexos e de alta dimensionalidade: economia humana, vida biológica e o clima de planetas inteiros, o nosso e outros.[95]

O JARDIM DA MEDICINA

A fragilidade da saúde humana tem sido responsável por mais mortes prematuras e sofrimento desnecessário do que qualquer conflito ou desastre natural de que haja memória.[96] Apesar de nos dois últimos séculos termos levado a cabo tentativas toscas e incompletas de decifrar e controlar o

código da vida, a nossa capacidade de proceder com maior precisão e propósito tem sido impedida por um único ingrediente em falta: nomeadamente, uma inteligência capaz de compreender esse código com um nível de pormenor suficiente. Agora, impulsionados pelo surgimento de uma inteligência que excede em muito a nossa, estamos a convergir para uma revolução na biologia que pode mudar a conceção que temos da vida humana.

Fundado no século XVII por decreto do rei Luís XIII de França e colocado sob a autoridade do médico real, o Jardin royal des plantes médicinales – o Jardim Real das Plantas Medicinais – foi o maior e mais avançado projeto do género na sua época. Expedições científicas a lugares tão distantes como Java e a Amazónia regressaram com uma variedade deslumbrante de plantas, meticulosamente estudadas quanto à sua potencial utilização medicinal por uma dedicada equipa de botânicos.

Menos de uma década mais tarde, o monarca francês sucumbiria à tuberculose; nenhum antídoto terreno tinha sido suficientemente forte para aliviar a sua dor ou atrasar a sua morte. Mas o jardim produziu com sucesso tratamentos e curas que não se encontravam em nenhum outro lugar do mundo.[97] Atualmente, a IA tem capacidade para ser uma biblioteca de produtos farmacêuticos igualmente fantástica, abrindo para benefício dos seres humanos um novo e vasto repositório de remédios para aliviar a doença e a tensão.

Extraordinariamente hábeis na geração de novas combinações de componentes selecionados e na identificação dos atributos com melhor desempenho de cada um, IA revolucionárias como a AlphaFold da DeepMind – com a sua gigantesca base de dados de mais de 200 milhões de previsões de estruturas proteicas – abriram novas perspetivas na saúde global. É provável que a IA produza avanços não só na engenharia de proteínas adicionais, incluindo novas hormonas, enzimas e anticorpos, mas também na identificação das causas moleculares de várias doenças e no desenvolvimento de potenciais tratamentos. Por sua vez, os cuidados médicos, com base na resolução sem precedentes da IA à escala molecular e genómica, poderão tornar-se cada vez mais personalizados – com medicamentos, e os respetivos métodos de administração, adaptados ao perfil metabólico único do paciente, ao risco de dependência, às tolerâncias estimadas e à suscetibilidade a possíveis efeitos secundários.

Os médicos empenhados no alívio do sofrimento humano teriam deste modo um parceiro para os ajudar a cumprir a sua vocação misericordiosa. As instruções da máquina poderiam guiar as mãos humanas mais hábeis – por exemplo, na neurocirurgia – em procedimentos há muito considerados possíveis mas demasiado arriscados no que respeita a segurança. A IA já está a ajudar a encontrar caminhos não destrutivos ou não intrusivos através do cérebro para remover mecanicamente uma fonte de debilitação ou curá-la biologicamente. Nos casos em que os problemas não são de natureza física mas psicológica, a IA poderá combater a degradação cognitiva, as doenças mentais, as perturbações psiquiátricas e, possivelmente, até a solidão.

Na realidade, e há muito a esperar desse lado, a IA poderia também afastar-nos do tratamento e aproximar-nos da *prevenção*, reduzindo assim a necessidade de curar. Atuando como mecanismos vigilantes de aviso precoce, os sistemas de IA poderiam alertar-nos para doenças malignas e anomalias muito antes de estas evoluírem para ameaças graves. Também ao nível social as IA poderão tornar-se sistemas avançados de monitorização da saúde, capazes de identificar e neutralizar doenças infeciosas antes de estas se transformarem em pandemias planetárias.

No entanto, todos os cenários acima referidos, mesmo o relacionado com a prevenção de doenças e mortes prematuras, são exemplos de remediação. São maneiras de a IA nos ajudar a atenuar ou a resolver problemas que atualmente arrastam a saúde individual para níveis inferiores aos padrões vigentes de bem-estar adequado. Mas o que dizer dos avanços que poderiam redefinir o *máximo* da saúde humana?

Ampliados pela IA, alguns progressos médicos evoluirão de terapias para extensões da longevidade humana. O recente fenómeno da edição de genes demonstra a proximidade e a viabilidade de tais progressos. Recorrendo a uma biotecnologia conhecida como CRISPR-Cas9 – e a uma outra variante chamada edição primária – os cientistas começam por identificar uma sequência genética específica que gostariam de manipular. Uma cadeia pré-concebida de ARN pode então guiar outra enzima especial até ao segmento escolhido de ADN nessa sequência e abri-lo para fazer as alterações e correções necessárias. Utilizando estas ferramentas, e as que lhes sucederem, poderá ser possível vencer não só as nossas deficiências como a própria mortalidade.

A morte, no entanto, sempre foi o limite divinamente imposto à humanidade, sem que nenhuma tentativa de isenção ficasse impune. Considere-se o mito de Sísifo, o astuto e ardiloso rei da antiga Éfira (atual Corinto), condenado a ficar aprisionado no submundo por ter irritado os deuses. Escapou à mortalidade graças a uma astúcia – conseguiu acorrentar Tânato, a personificação da morte, com as suas próprias grilhetas antes de fugir para regressar à terra dos vivos. Mas sem a morte para levar as pessoas (ou as coisas), desencadeou-se o inferno na Terra. Os velhos e os doentes sofriam sem esperança de um fim, o gado não podia ser abatido para consumo e não era possível matar animais para os sacrificar aos deuses.

A crise foi finalmente resolvida quando Ares, o deus grego da guerra, interveio, libertando Tânato e permitindo que Sísifo escapasse pela segunda vez do fundo abismo do Tártaro. Mas a brincadeira acabou ali: os deuses estavam fartos e fizeram do rei grego um exemplo e um sério aviso antes que acontecesse uma terceira transgressão.

A morte continua a ser o grande igualizador da vida. Até o imperador chinês Qin Shi Huang[98] – que se dizia ser filho do céu e foi um dos primeiros a procurar o elixir da vida – morreu por ter bebido demasiado mercúrio e foi sepultado juntamente com o seu grande exército de terracota. Por vezes, procurar a imortalidade parece só servir para apressar o seu oposto.

Além disso, a nossa impermanência tem as suas vantagens. Concentra a mente e confere uma maior urgência aos nossos esforços. Como o escritor americano Jack London declara triunfantemente: «A função do homem é viver, não existir. Não desperdiçarei os meus dias a tentar prolongá-los. Usarei o meu tempo.»[99] Tolstoi escreve (citando Sócrates), «só nos aproximamos da verdade na medida em que nos afastamos da vida».[100] Quando estava a morrer de um cancro em fase terminal, o grande físico John von Neumann pediu que um padre católico o acompanhasse nos seus últimos dias – a sua fé deixara de estar em conflito com a ciência agnóstica de que tinha sido pioneiro.[101] Não haverá usos para os finais?

Hoje ou amanhã, podemos dar por nós na situação de ter de verificar até que ponto a IA pode prolongar as nossas vidas. Uma redução da nossa consciência da mortalidade iminente poderia, aliás, fazer mudanças incalculáveis na psique humana. A seu tempo, as sociedades poderão

precisar de decidir coletivamente a duração ideal de uma vida humana e, ao fazê-lo, responder às questões metafísicas e espirituais que uma tal decisão acarreta: será a longevidade humana meramente o produto de uma expetativa comunitária, um limite que pode ser posto de lado como uma falsa fronteira autoimposta ao potencial da nossa espécie? Ou deveríamos considerar o tempo de vida humano, seja ele natural ou divinamente outorgado, uma restrição sagrada ao poder de uma dada pessoa? Estas questões são mais profundas do que qualquer procura individual de uma biologia ótima.

Mesmo que a duração total da vida não fosse alterada, no futuro poderíamos talvez garantir que ela não terminaria num declínio prematuro nascido da vulnerabilidade biológica. Mas ir demasiado longe na eliminação dessa fraqueza pode também ter efeitos secundários. Não foi o facto de triunfarmos sobre os desafios, incluindo as doenças, que nos tornou respeitáveis? Por outro lado, mesmo que curássemos todas as doenças ou nos «fabricássemos» de modo a tornarmo-nos imunes a elas, os seres humanos continuariam a ser vulneráveis de outras maneiras: um acidente físico, a ruína financeira, o desgosto.

Numa visita a Paris, em 1833, Ralph Waldo Emerson ficou fascinado com o Jardim Real das Plantas Medicinais, atónito ao ver quanto mais avançado o mundo natural parecia – em toda a sua variedade e forma – do que as máquinas da Revolução Industrial.[102] Desorientado numa era tecnologicamente caótica, o estado de espírito transatlântico era uma mistura ambígua de admiração e alarme. Ao regressar desse «célebre repositório de curiosidades naturais», Emerson retirou-se para as zonas rurais do Massachusetts e, inspirado pelo que vira, articulou uma resposta humana que voltou a situar o mundo natural no centro do mundo agora mecanizado. O seu diagnóstico realçava a maneira como a biologia podia guiar-nos através de tempos desconhecidos, servindo ao mesmo tempo como o nosso melhor lembrete de que os seres humanos continuam a ser os derradeiros «definidores e criadores de mapas das latitudes e longitudes da nossa condição».[103]

Meio século após a visita de Emerson, e dois séculos e meio após a sua fundação, tendo passado por períodos de revolução violenta e de revisão curatorial, o jardim transformou-se para lá das necessidades imediatas do

seu primeiro patrono para se tornar um museu para a história do que era então uma ideia nova e controversa. O seu nome: *La galerie de l'Évolution*, a Galeria da Evolução.[104]

A evolução está prestes a ser redefinida na era da IA, uma vez que certas ferramentas levantam a perspetiva da autoengenharia humana. Por exemplo, a maior parte da edição de genes que os instrumentos modernos possibilitam limita-se a células somáticas ou não produtivas. No entanto, é possível fazer alguma edição em células germinativas, cujas características são herdadas reprodutivamente. É possível que algumas pessoas decidam «corrigir» doenças congénitas nos seus descendentes. Outras poderiam ir mais longe, optando por instalar vantagens congénitas – vantagens que poderão não ser possibilitadas pelo genoma de qualquer dos pais biológicos ou, no extremo, o de qualquer outro ser humano. Isso iria para lá da elevação da raça humana; seria verdadeiramente redesenhá-la.

É possível que em breve venhamos a ter o poder de determinar o ritmo e a direção da nossa própria espécie. Esta ideia é tão controversa hoje como a evolução o era no tempo de Emerson, e a perspetiva levanta a questão óbvia mas insidiosa: como é o ser humano perfeito? Esta pergunta tem sido feita e respondida por várias sociedades e, em alguns casos, serviu de base a empreendimentos «científicos» e políticos que provocaram grandes tragédias humanas. Por isso convém também perguntar, com alguma apreensão: devemos tentar descobrir?

Talvez estas experiências sejam um sacrilégio. Ou talvez a capacidade humana de inventar estas tecnologias seja, por si só, um indício de que o que entendemos como nosso limite sempre esteve destinado a ser ultrapassado. Se existe um Criador, teremos sido criados para podermos criar-nos a nós mesmos? Em caso afirmativo, é nosso dever garantir a manutenção da nossa capacidade de ação humana na era da IA? Diferentes comunidades darão respostas diferentes a estas perguntas. Nenhuma conseguirá escapar à necessidade urgente de uma resposta.

ENGENHARIA PLANETÁRIA

A história da Terra é tão violenta como incompreendida. Existe uma janela muito estreita de condições climáticas adequadas para o florescimento da vida. Um pouco mais frio – como já aconteceu em nada menos que cinco eras glaciais – e a nossa Terra torna-se uma rocha desolada e congelada.[105] Um pouco mais quente – como está a acontecer agora – e torna-se um inferno abrasador. Em *Os Irmãos Karamazov*, de Fyodor Dostoevsky, o demónio diz a Ivan:

> A Terra na qual vivemos pode ter sido repetida um bilião de vezes. Foi extinta, congelada, rachada, partida em pedaços, desintegrada nos seus elementos, novamente «as águas por cima do firmamento» [Génesis 1:7], novamente um cometa, novamente um sol, novamente terra a partir de um sol – e a mesma sequência pode ter sido repetida infinitamente e exatamente igual em todos os pormenores, o que é indecente e insuportavelmente entediante.[106]

Embora possamos não partilhar a exasperação do anjo caído, estamos conscientes da ciclicidade da história geológica do nosso planeta. Todas as diversas causas das cinco anteriores extinções em massa – desde o impacto instantâneo de meteoros à formação gradual de glaciares – resultaram das vicissitudes extremas do clima terrestre.[107]

O problema atual é, evidentemente, uma aceleração decisiva em direção a um novo extremo de calor. Trata-se, na realidade, de dois problemas distintos, ambos derivados de uma dependência excessiva do carbono. Nós, os autores deste livro, estamos convencidos (talvez com muito otimismo à mistura) de que ambos são causados pela química, e em última análise resolúveis pela química – se for possível desbloquear todas as suas capacidades utilizando a IA.

O primeiro problema é o aquecimento atmosférico. Os nossos males atuais foram provocados pela transferência demasiado rápida de carbono fóssil da geosfera subterrânea para a biosfera ativa acima do solo.[108] As tentativas de engenharia dos sistemas climáticos – de que existem duas aplicações principais – sempre foram difíceis na teoria e ainda mais difíceis de testar na prática. Mas onde nós somos travados, a IA pode não o ser. Uma das aplicações é a remoção de carbono, que inverte a transferência do excesso deste elemento – no caso vertente da atmosfera para o subsolo, a

geosfera. A principal solução neste domínio baseia-se numa química simples descoberta há mais de duas décadas. A IA poderá identificar um método novo e mais eficiente.[109]

A outra é a geoengenharia solar: a libertação de certas partículas na atmosfera para refletir a luz do Sol e «arrefecer a Terra». A engenharia solar, tal como a remoção do carbono, poderia reduzir os efeitos das alterações climáticas, permitindo-nos evitar algumas das consequências mais extremas da acumulação de carbono. (Nenhuma delas abordaria, porém, as causas radiculares das alterações climáticas.) Outras ideias hipotéticas para a engenharia dos céus baseiam-se em jatos de alta altitude e em elementos projetados por supervulcões. Se alguma vez forem experimentadas em grande escala, poderão, como a remoção do carbono, revelar-se estratégias toscas e arriscadas.

Uma IA, no entanto, se concebida para integrar dados de instrumentos em terra, no mar e no espaço, poderia criar um modelo extremamente pormenorizado e em tempo real do clima da Terra. A um nível muito elevado de granularidade, a química atmosférica do nosso planeta poderá mostrar-se não como o sistema caótico que vemos, mas como apenas mais um processo industrial de entradas e saídas químicas exatas, todas passíveis de uma gestão precisa.

A IA pode também responder a incidentes pontuais que ameacem o precário equilíbrio climático da Terra, seja a erupção de um supervulcão que projete matéria para a atmosfera superior ou a detonação de bombas atómicas e a sua ameaça de um inverno nuclear – ainda que neste caso possa apenas evitar o colapso ecológico sem nada fazer quanto à catástrofe humana imediata. Mesmo assim, ao intervir à escala planetária, uma IA habituada ao controlo preciso do clima terrestre estaria bem posicionada para manter a nossa casa habitável.

Acreditamos, portanto, que a IA oferece uma fonte de esperança numa luta que é demasiadas vezes retratada como uma guerra que não podemos ganhar. No entanto, mesmo que as IA se revelem capazes destas intervenções, seria perigoso ficar demasiado dependente delas. Essa ajuda deve ser considerada um complemento e nunca um substituto.

Além disso, o segundo problema – o problema da energia, distinto do problema do aquecimento atmosférico – continua a exigir uma solução separada. Uma vez que os hidrocarbonetos, a nossa principal fonte de

energia, demoram milhões de anos a formar-se e são consumidos num punhado de séculos, os seres humanos veem-se obrigados a encontrar uma alternativa para a energia planetária, independentemente de quaisquer alterações na composição atmosférica.[110]

Se a nossa suposição de que a IA pode afinar a química da nossa atmosfera estiver de um modo geral correta, é verosímil que seja capaz de fazer o mesmo em relação aos nossos produtos energéticos, concebendo e produzindo novos bens isentos de carbono para substituir os seus problemáticos antecessores. Poderíamos imaginar uma IA a testar e combinar milhões de substitutos sintéticos numa simulação virtual ou num laboratório físico até chegar a substitutos de carbono zero para o petróleo, o gás e o carvão.

Idealmente, a IA poderia também criar microrganismos otimizados para produzir estes novos combustíveis – e poderia conceber os processos de utilização desses novos produtos de modo a adaptarem-se às refinarias existentes e funcionarem com equipamento antigo. Ao gerar energia nos mesmos locais, utilizando a mesma maquinaria, mas não através dos mesmos métodos, estes substitutos sustentáveis poderiam ser compatíveis não só com as infraestruturas da sua produção mas também com as do seu transporte e consumo. Outra possibilidade ainda é que a IA venha finalmente a desbloquear a viabilidade da fusão nuclear a frio, o que resolveria não só o problema energético do nosso planeta, mas também o de qualquer outro planeta onde possamos um dia encontrar-nos.

Alguns vão recusar estas ideias, considerando-as irrealistas (especialmente tendo em conta a urgência do nosso desafio imediato) ou talvez indesejáveis – ou seja, contrárias ao nosso desejo de deixar a natureza inalterada, semelhante ao instinto que tão fortemente motiva os nossos esforços para a preservar. Como adverte o ambientalista indígena equatoriano Nemonte Nenquimo, «A Terra não espera que a salvemos, espera que a respeitemos.»[111]

Acreditamos que o nosso otimismo está conforme às exigências de respeito e humildade. Também admitimos que a única coisa que sabemos é que o futuro é incognoscível. De modo algum sugerimos ou insinuamos que é necessária uma superinteligência para nos tirar da situação difícil que nós próprios criámos. Os seres humanos são perfeitamente capazes de o fazer mesmo sem a ajuda da inteligência artificial, e não devemos parar ou

abrandar os nossos esforços atuais. Pelo contrário, devemos acelerá-los. Isto é especialmente verdade quando reconhecemos a enorme quantidade de energia que provavelmente será consumida pelo treino e pela inferência das IA – procedimentos que, se não forem acompanhados por nenhum dos esforços intencionais e direcionados que aqui descrevemos, só virão agravar uma situação já de si péssima.

Ao mesmo tempo, se o silício puder efetivamente fornecer soluções adicionais para os problemas criados pelo carbono, seria sensato explorarmos essa oportunidade. Isto é particularmente verdade dados os enormes benefícios potenciais para o mundo desenvolvido, o mais afetado pelos problemas decorrentes das alterações climáticas – e também o que sofreria na eventualidade de ser imposto um limite mundial ao consumo de energia.

A IA é muito frequentemente vista como mais um produto da mesma filosofia destrutiva responsável pelo nosso atual dilema climático. Preocupa-nos esta visão ser demasiado míope e poder fazer-nos perder uma oportunidade significativa de corrigir o nosso presente sem recuar para um passado pré-industrial. Esperamos, tal como Winston Churchill em 1940, ser testemunhas de que «o Novo Mundo, com todo o seu poder e força, avança para salvar e libertar o Velho».[112]

MAIS ALÉM DO NOSSO PLANETA

Enterrada nas profundezas das montanhas do sudoeste da China, a humanidade suspende a respiração, a escutar o bater do seu próprio coração enquanto espera pacientemente que os primeiros gritos surjam do mar de silêncio cósmico. Este lugar remoto alberga o maior radiotelescópio da Terra – um gigantesco prato metálico liso, tão grande como os picos que o escondem – uma máquina apelidada «Olho do Céu» e construída à medida para procurar vida para lá do nosso planeta. Mas embora a vastidão e antiguidade do Universo signifiquem que deveria haver uma cacofonia de sinais de outras civilizações – ou, pelo menos, dos restos de civilizações anteriores –, o que há em vez disso é um silêncio quase absoluto.

Pensamos que sabemos o que procurar. Tentamos adivinhar as tecnologias que uma civilização suficientemente avançada poderá ter desenvolvido e prever quais dos seus sinais poderemos plausivelmente ver ou ouvir. Quando procuramos *flashes* de energia atómica na atmosfera de mundos distantes ou sondamos densos campos estelares em busca de estruturas físicas anormalmente grandes, estamos limitados pelas capacidades dos instrumentos de que dispomos – incluindo a nossa imaginação.

Se existir vida além da nossa, os astrónomos chineses deste posto avançado poderão ser os primeiros a ouvi-la.[113] Agora, porém, perguntamo-nos se será assim tão provável que sejamos nós a estabelecer o primeiro contacto. Talvez esse primeiro contacto seja feito por outra inteligência alienígena – nascida no nosso planeta.

A IA já está a ajudar-nos a escutar e a procurar vida extraterrestre, peneirando milhares de milhões de assinaturas tecnológicas mais antigas, separando as que foram causadas por interferência humana das originadas por fontes potencialmente exteriores.[114] Onde os humanos só ouviriam estática, a IA pode reconhecer comunicações até agora indecifráveis ou ignoradas.

Galileu descreveu certa vez o nosso universo como um «grande livro» escrito na linguagem da matemática. Se as formas de vida extraterrestre aprenderam este método universal e preciso de representação e raciocínio, tal como nós, ou se as suas línguas, tal como a nossa, puderem ser traduzidas para essa linguagem, poderemos ser capazes de interpretar – e responder – aos seus sinais. O mais provável é que este esforço seja pelo menos assistido, se não mesmo conduzido, pela IA. As IA poderiam tornar-se muito mais do que observadores e recetores passivos – não apenas tradutores, mas também aventureiros, sinalizadores e batedores. Poderiam servir como astronautas, indo mais longe do que até a imaginação humana é capaz de alcançar. Os humanos do futuro poderão inclusivamente acompanhar as suas IA para lá do sistema solar. Juntos, poderão iluminar a existência de civilizações há muito desaparecidas e descobrir as razões da sua extinção – talvez educando-nos sobre possíveis perigos futuros. As IA familiarizadas com o espaço profundo poderiam descobrir novos e abundantes materiais orgânicos. Num futuro distante, poderiam ajudar na engenharia de megaestruturas planetárias para proteger a Terra do impacto

de cometas e asteroides ou de ser tragada por buracos negros. Sistemas de IA poderiam transformar atmosferas planetárias de modo a adaptá-las a nós, ou ajudar-nos a ajustar a nossa fisiologia de modo a tornar mais fácil a aclimatação a outro sistema.

É claro que a IA pode também tornar-se a razão do fracasso da nossa civilização – acabando com a humanidade ao pôr-nos em contacto com outra inteligência alienígena, mas hostil. Por exemplo, a IA poderia possibilitar a emissão de sinais mais altos e mais duradouros para o espaço, revelando a presença da civilização humana a quem quer que estivesse a ouvir no escuro. Também poderia procurar outros mundos habitados, aumentando grandemente a probabilidade de encontrar um – com consequências imprevisíveis. Há quem argumente que, se quisermos entrar na arena da exploração espacial, só devemos fazê-lo depois de atingirmos um nível de sofisticação tecnológica que nos permita defendermo-nos de qualquer inimigo imaginável – sugerindo que a IA é um pré-requisito, e uma salvaguarda posterior, de qualquer encontro com vida extraterrestre. Mas se encontrarmos uma espécie inteligente pós-biológica, os nossos parceiros de silício não serão garantia de uma receção de braços abertos.

A outra maneira de a IA poder significar o fim da civilização humana é, evidentemente, desalinhando-se da humanidade. (Essa possibilidade pareceria ainda mais premente se descobríssemos provas de uma extinção de vida biológica provocada por uma máquina noutro planeta). Durante eras, o fascínio pela vida além da Terra alimentou os nossos escritos, animou as nossas religiões e ocupou os nossos astrónomos. Estamos convencidos de que a humanidade pensou durante mais tempo e mais aprofundadamente sobre mundos distantes do que sobre essa outra possibilidade mais provável – um «primeiro contacto», aqui na Terra, com criaturas que nós próprios criámos. Devia preocupar-nos tanto – se não mais – a necessidade de compreender e dar forma à natureza e intenção dos que estão aqui connosco como às dos que estão longe no espaço e no tempo – um tema que será objeto de uma ampla discussão no próximo capítulo.

Mesmo que equipas de humanos e IA não descubram formas de vida extraterrestre ou não abram caminho para planetas extraterrestres, poderão obter conhecimentos úteis e até então alienígenas. Juntos, poderíamos chegar a uma compreensão mais elevada do âmbito do universo, da natureza do espaço-tempo, da estabilidade dos sistemas estelares e da

natureza da gravidade na astrofísica. A modelação mecânica da IA poderia encolher as distâncias estelares, comprimir e expandir o tempo humano na Terra e transformar e alargar a nossa perceção de todos os limites e aspetos do universo.

O estudo da astrofísica com recurso à IA poderia permitir uma compreensão humana mais profunda da nossa origem num tempo impensavelmente antigo. Há muito que a cosmologia religiosa procura dar respostas filosóficas ou teológicas a estas questões das origens do homem. O que encontrarmos na próxima fase de exploração universal – e que implicará necessariamente o nosso passado mais profundo – poderá colorir a maneira como percebemos algumas das crenças mais sagradas da humanidade. Talvez haja uma razão para que a relíquia mais sagrada do islão seja a Pedra Negra, um meteorito que se encontra no interior da Casa de Deus em Meca, o centro do ritual no mundo muçulmano.[115] Talvez haja mais do que parece no conceito cabalístico de *kefitzat haderekh*, a expressão hebraica para designar um atalho milagroso (literalmente, «contração da estrada») entre dois lugares distantes num curto espaço de tempo.[116]

Em *The Grand Design*, Stephen Hawking e o seu coautor Leonard Mlodinow escrevem: «O passado (não observado), tal como o futuro, é indefinido e existe apenas como um espectro de possibilidades.»[117] Por enquanto, é esta a perceção humana. Resta saber se inteligências mecânicas tão precisas como as que estamos a desenvolver agora poderão perturbar esta incerteza, identificando apenas uma realidade. As nossas criações podem, assim, transmitir uma única história universal, uma história que parece simultaneamente vulgar e especial, impossivelmente pequena e milagrosamente divina.

---

95 Ver Donella H. Meadows, *Thinking in Systems* (White River Junction, VT: Chelsea Green Publishing, 2008).

96 O número de pessoas que morrem naturalmente de velhice continua hoje a ser superior ao das que morrem de causas não naturais decorrentes de problemas de saúde.

97 Antoine de Jussieu levou plantas de café de Java para Paris. Ver Deligeorges *et al.*, *Le Jardin des Plantes et le Muséum National d'Histoire Naturelle* (Paris: Patrimoine, 2004), 13-15. O Jardin des Plantes foi originalmente criado em 1635 pelos médicos do rei Luís XIII para albergar o herbanário medicinal do rei e era por eles supervisionado em nome de Sua Majestade Real. O rei Luís XIII morreu a 14 de maio de 1643, com 41 anos. Para a expedição científica à Amazónia, ver a Missão Geodésica Francesa ao Equador no século XVIII.

98 Ver Sima Qian, *Records of the Grand Historian: Han Dynasty II* (Nova Iorque: Columbia University Press, 1993).

99 Jack London, citado pelo seu testamentário literário, Irving Shepard, numa introdução a uma coleção de histórias de London editada em 1965. Jack London, *Jack London's Tales of Adventure*, ed. Irving Shepard (Springdale, AR: Hanover House, 1956), VII.

100 Tolstoi cita Sócrates em *A Confession and Other Religious Writings*, trad. David Patterson (Nova Iorque: W. W. Norton, 1983), 43: «"Só nos aproximamos da verdade na medida em que nos afastamos da vida", diz Sócrates, enquanto se prepara para a morte. "Que procuramos na vida, nós os que amamos a verdade? Sermos livres do corpo e de todos os males que resultam da vida do corpo. Se assim é, como podemos deixar de nos regozijar quando a morte se aproxima?"» Sócrates discute este assunto nas secções 62-69 do *Fédon* de Platão, quando os amigos vão vê-lo uma última vez antes da execução.

101 As crenças religiosas de Von Neumann (ou a falta delas) têm sido objeto de muita discussão. Judeu de nascimento, aceitou o batismo católico em 1930 para se poder casar, embora não praticasse a fé e alguns dos seus colegas o considerassem «completamente agnóstico». Por isso foi uma grande surpresa para todos quando, a morrer de cancro no hospital, procurou os préstimos de um sacerdote católico, o padre beneditino Anselm Strittmatter, a quem se confessou e de quem recebeu os últimos sacramentos da Igreja Católica. Sociedade dos Cientistas Católicos, Cientista Católico do Passado: John von Neumann, https://catholicscientists.org/scientists-of-the-past/john-von-neumann.

102 Maurice York e Rick Spaulding, *Ralph Waldo Emerson: The Infinitude of the Private Man* (2008); Robert D. Richardson, *Emerson: The Mind on Fire* (1995); Ronald Bosco e Joel Myerson, *The Selected Lectures of Ralph Waldo Emerson* (2005).

103 Ralph Waldo Emerson, *The Complete Works of Ralph Waldo Emerson* (Boston: Houghton Mifflin, 1904), vol. 4, n.º 12; palestra de Ralph Waldo Emerson, «The Uses of Natural History» perante a Boston Natural History Society no Masonic Temple em Boston, 5 de novembro de 1833. Posteriormente revista, melhorada e editada no seu primeiro livro, *Nature*, em 1836.

104 O Jardin royal des plantes médicinales foi fundado em 1635; a Revolução Francesa começou em 1789; a Galeria da Evolução foi inaugurada em 1889. A ideia de evolução (mas não a seleção natural como seu mecanismo) foi referida por 70 autores diferentes entre 1748 e 1859, o ano em que Darwin publicou *A Origem das Espécies*.

105 O termo «idade do gelo» pode ser enganador porque as idades do gelo são tecnicamente classificadas como uma mistura de avanço dos glaciares (período glacial) ou recuo dos mesmos (período interglacial). Embora os períodos interglaciais sejam relativamente quentes, continuam a ser classificados como parte de uma época glaciar. A época atual continua a ser tecnicamente classificada como uma idade do gelo, uma vez que estamos num período interglacial.

106 Fyodor Dostoevsky, *The Brothers Karamazov*, trad. Constance Garnett (Nova Iorque: The Modern Library, 1900), 783.

107 Depende de como se conta; outras estimativas afirmam que a Terra conheceu 20 extinções maciças. Algumas são ainda hoje objeto de debate.

108 Conferência dada pelo Dr. David Keith na Gustavus Adolphus College Nobel Conference: «How Might Solar Geoengineering Fit into Sound Climate Policy», 25 de setembro de 2019, https://www.youtube.com/watch?v=Ia1AWdmRsMc&t=234s.

109 O *looping* de cálcio, um processo proposto pela primeira vez em 1999 por químicos japoneses, utilizado em tecnologias de captura direta de ar. Ver Shimizu, Hirama e Hosoda, «A Twin Fluid-Bed Reator for Removal of CO2 from Combustion Processes», *Chemical Engineering Research and Design*, 77, n.º 1, janeiro de 1999, 62-68.

110 As estimativas são complexas e imprecisas, mas a maior parte das análises sobre esta questão produz prazos que diferem apenas em décadas. Ver várias análises efetuadas pela U.S. Energy Information Administration, Stanford, BP, e pelo Energy Institute. Poucas fontes fiáveis listam reservas de petróleo e gás com base nas taxas de produção e consumo atuais e projetadas que durem mais de 100 anos, 200 no caso do carvão.

111 Nemonte Nenquimo, «This is my message to the western world – your civilisation is killing life on Earth», *The Guardian*, 12 de outubro de 2020. Nenquimo é membro da Nação Waorani da Região Amazónica do Equador.

112 Discurso de Winston Churchill «Lutaremos nas praias», proferido na Câmara dos Comuns a 4 de junho de 1940: «O novo mundo, com todo o seu poder e força, avança para salvar e libertar o Velho.» Ver *Never Give In! The Best of Winston Churchill's Speeches* (Londres: Pimlico, 2004), 218.

113 Ver Ross Andersen, «What Happens If China Makes First Contact?», *The Atlantic*, 15 de dezembro de 2017.

114 Peter Ma *et al.*, «A deep-learning search for technosignatures from 820 nearby stars», *Nature Astronomy* 7, 492-502, January 30, 2023.

115 Referência à Kaaba, que alberga uma pedra negra caída do céu para ligar o firmamento à terra, exposta no centro do Masjid al-Haram em Meca, Arábia Saudita.

116 Yeshaya Elazar, «Kefitzat Haderech: What's the Message of This Rare Form of Divine Intervention?» Chizuk Shaya (blog), 29 de novembro de 2009. https://www.chizukshaya.com/2009/11/kefitzat-haderech.html.

117 Stephen Hawking e Leonard Mlodinow, *The Grand Design* (Nova Iorque: Bantam, 2010), 82.

PARTE III

A ÁRVORE DA VIDA

CAPÍTULO 8

ESTRATÉGIA

No século XX, a história forçou as sociedades humanas a empenhar-se numa série de projetos monumentais com destaque para os choques fulminantes das duas guerras mundiais e o consequente desenvolvimento de uma arquitetura internacional para evitar a sua repetição; a lenta decadência dos impérios e a organização de Estados independentes para facilitar a reconstrução pós-colonial; a rápida expansão das forças comerciais e tecnológicas, e a reafirmação da autonomia – individual, cultural e nacional – para moderar o avanço dessas forças.

De muitas maneiras, a humanidade terminou o século mais pacífica, igualitária e ligada do que nunca. Noutros aspetos, porém, os nossos esforços coletivos falharam: o sofrimento humano básico, as desigualdades globais e a possibilidade de confrontos cataclísmicos entre rivais geopolíticos continuam presentes e persistentes.

Além disso, enfrentamos hoje um desafio mais complexo, mais existencial e de uma natureza muito diferente dos anteriores, com a desvantagem de não dispormos de décadas, quanto mais de um século, para o enfrentar. As escalas temporais comprimidas pelas IA – de que já falámos no Capítulo 2 – dão-nos menos margem de manobra para agir, e uma maior tolerância da nossa parte poderia resultar em catástrofe. Numa janela de tempo perigosamente estreita, os nossos esforços coletivos têm de ser ainda mais bem-sucedidos do que as realizações do século passado.

Um mundo em turbulência apresenta, no entanto, muitas possibilidades de ação produtiva, que as mais das vezes requer apenas decisões táticas. Num momento como este, a manobra mais eficiente e eficaz consiste em definir os fundamentos de uma estratégia que possa orientar as escolhas tanto deste dia como de todos os dias previsíveis. A articulação de princípios estratégicos pode estabelecer limites úteis para o que é concebível, fornecer bases para decisões isoladas e diminuir o peso mental quando a crise inevitavelmente chegar.

A nosso ver, há uma questão que deve definir a estratégia humana nesta nova era de ajuste de contas, que é a seguinte: vamos nós tornar-nos mais parecidos com elas (as IA), ou elas mais parecidas connosco? É uma questão que já foi levantada muito cedo neste livro, praticamente nos mesmos termos. Responder-lhe continua a ser a nossa primeira e mais necessária tarefa.

Este capítulo tenta dar uma resposta preliminar. Ao fazê-lo, discute várias ideias grandes, problemáticas e que talvez pareçam vagas, desde a «coevolução» do orgânico (i.e., o biológico) e do sintético (i.e., o artificial) até à natureza da segurança inteligente e da inteligência segura, passando pela definição de humano. Embora as ideias em si não tenham sido ditadas do alto nem gravadas na pedra, as suas várias implicações para os atores futuros representam um desafio formidável. No entanto, apesar da dificuldade, não nos podemos esquivar ao trabalho filosófico, técnico e diplomático necessário para compreender essas implicações e dar e aplicar na prática respostas sensatas. Ao assumirmos este fardo, podemos estar gratos por não ser demasiado tarde para trabalharmos agora no sentido de garantir que as boas intenções humanas vão prevalecer para lá deste ponto de charneira da história.

COEVOLUÇÃO: HUMANOS ARTIFICIAIS

Até agora, a história da computação tem seguido uma trajetória de integração e interação cada vez maiores entre humanos e máquinas. Enquanto moldávamos as nossas ferramentas para se adaptarem cada vez

mais a nós – de acordo com milénios de prática humana – nunca considerámos o desenvolvimento de outras que fossem inadequadas à nossa anatomia ou ao nosso intelecto, guiando-nos única e exclusivamente pelos limites da nossa biologia. Agora, porém, o advento da IA pode persuadir pelo menos alguns de nós a ponderar a possibilidade de uma missão inversa: no caso de as ferramentas que moldamos parecerem ultrapassar as capacidades que possuímos – como a IA por vezes já faz – seremos capazes de considerar a hipótese de nos refabricarmos de modo a maximizar a utilidade das ferramentas e assim garantir a nossa participação contínua em empreendimentos partilhados como os descritos nos capítulos anteriores?

Estão já em curso esforços de engenharia biológica destinados a conseguir uma fusão mais estreita do homem com a máquina. Começando com interligações físicas através da implantação de *chips* no cérebro humano[118], esses esforços procuram uma maneira mais rápida e eficiente de estabelecer uma ponte entre a inteligência biológica e a inteligência digital. A criação destas ligações poderia aumentar a nossa capacidade de comunicar com as máquinas, desafiá-las nos seus próprios termos, garantir que o conhecimento adquirido pela IA é transmitido aos humanos e convencer a IA do valor desses humanos como parceiros iguais.

Na realidade, as tentativas de construir «interfaces cérebro-computador» não só podem reforçar o esforço da humanidade para se integrar com as máquinas, como a engenharia neuronal pode ser apenas uma fase intermédia de transição para uma simbiose real. Alcançar a verdadeira paridade com a IA exigiria provavelmente passos que vão para lá da modificação individual. Por exemplo, uma sociedade pode tentar conceber uma linha genética hereditária personalizada para ser compatível com a IA. Essas novas interligações entre a inteligência biológica e a inteligência artificial poderiam contornar, ou remeter para o passado, as ineficiências humanas na absorção e transmissão de conhecimentos.

Mas os perigos – éticos, físicos e psicológicos – de um tal percurso podem facilmente ultrapassar os benefícios. Se conseguirmos rever a nossa biologia (provavelmente através da utilização de IA), corremos o risco de perder uma linha de base para fundamentar o nosso pensamento futuro sobre as possibilidades ou os perigos que poderemos enfrentar enquanto espécie. Por outro lado, se não adquirirmos essas novas capacidades podemos colocar-nos numa situação de desvantagem na coexistência com a

nossa criação. Tal como as coisas estão de momento, uma autorremodelação extrema pode não ser necessária – e na realidade nós, os autores, consideramo-la indesejável. Mas a escolha entre alternativas que agora parecem fantasiosas poderá em breve ter de ser confrontada como real.

Entretanto, ao tentarmos navegar pelo nosso papel quando já não formos os únicos ou sequer os principais atores no planeta, poderemos alargar o nosso pensamento com um olhar para a história da própria coevolução biológica. Charles Darwin escreveu longamente sobre o curioso processo pelo qual as espécies afetam a evolução umas das outras.[119] Apesar de nunca ter usado a palavra nos seus escritos, foi dos primeiros a reconhecer que a coevolução é uma força importante de organização da vida na Terra.

Os genomas das espécies que interagem estão ligados; mudam em resposta uns aos outros ao longo do tempo. Os bicos longos e finos dos colibris, por exemplo, e os compridos funis da corola de certas flores cresceram juntos até dimensões mais extremas para servir as necessidades mútuas. Enquanto os líderes religiosos da época de Darwin acreditavam que tais adaptações personalizadas eram um sinal do desígnio divino, Darwin apresentou provas de outra explicação.

E a coevolução pode não ser exclusiva das espécies terrenas. Na astrofísica, uma teoria propõe que toda a expansão do cosmos pode ser atribuída à coevolução, com os buracos negros e as galáxias a desenvolverem-se de uma maneira interdependente não muito diferente da dos colibris e das flores.[120] Além disso, no sentido em que a coevolução envolve múltiplas partes que concebem novas disposições internas em resposta umas às outras, encontramo-la do mesmo modo no casamento entre pessoas, nas plataformas dos partidos políticos e nas relações entre nações – como, por exemplo, nas evoluções ofensivas e defensivas que em última análise estabilizaram a dinâmica nuclear durante a Guerra Fria.

Talvez a coevolução seja, nesse caso, a regra, e a estase a exceção? Se assim for, há que perguntar se a ausência de mudanças na espécie humana, até agora, apesar do nascimento da IA, é ela própria um desenvolvimento natural. E se não for, qual deve ser a nossa resposta? Devemos procurar a todo o custo um progresso humano acelerado, quer por lealdade ao conceito de evolução, quer por receio da alternativa?

Alguns temem que, com a chegada de uma tecnologia dotada de inteligência «superior», estejamos a enfrentar a nossa própria extinção. Que fazer? Se essa possibilidade não é mais do que um efeito colateral lógico da coevolução a seguir o seu curso, devemos ou não revoltar-nos? Como diz o filósofo francês Alain Badiou, «é o próprio mar que modela os barcos, escolhendo os que funcionam e destruindo os outros».[121] Para sobreviver teríamos então de aprender, como no passado, a construir barcos melhores. Neste cenário, a IA funciona primeiro como a nossa principal ameaça e depois, idealmente, como nossa parceira.

Se, no entanto, adotarmos esta abordagem, ao tentarmos mitigar os riscos de uma tecnologia estaremos, paradoxalmente, a aumentar os riscos de outra. Biologicamente – ou, pior, geneticamente – algo poderia correr mal. A especiação poderia fazer que a raça humana se dividisse em várias linhas, algumas infinitamente mais poderosas do que outras. Se em alguns casos a diferença seria desejável – por exemplo, na criação de um grupo de humanos biologicamente concebidos para o espaço – noutros poderia reforçar ainda mais as desigualdades ao longo das linhas de fratura existentes dentro das sociedades humanas e no seio delas.

Alterar o código genético de alguns humanos para os tornar sobre-humanos acarreta outros riscos morais e evolutivos. Se a própria IA for responsável pelo aumento da capacidade mental humana, poderá criar na humanidade uma dependência biológica e psicológica simultânea da inteligência «alienígena». Não é claro como poderiam os seres humanos, após uma íntima ligação física e intelectual, ultrapassar facilmente essa dependência de modo a desafiar as máquinas ou divorciarmo-nos delas, se necessário. Tal como tem acontecido com outras tecnologias, a adoção e a integração podem resultar numa dependência difícil de desemaranhar.

Talvez o mais preocupante seja a nossa ignorância coletiva: podemos nem sequer perceber que nos fundimos. E se nos apercebêssemos, seria uma pessoa vulgar capaz de reconhecer ou identificar um defeito – ou uma deserção – num humano com capacidades semelhantes às de uma máquina? Suponhamos que as preocupações com a segurança poderiam ser demonstravelmente dissipadas. Mesmo assim, a mudança de mentalidade decorrente da autorremodelação da humanidade ao serviço de uma parceria íntima ou a dependência de ferramentas à base de silício continuaria a ser um desenvolvimento extremo. Voltando a citar Tolstoi: «Sem controlo sobre

a direção, há menos consideração pelo destino.»[122] Para onde quer que a tecnologia nos leve, é para lá que iremos, a bem ou a mal. Ou, como já foi observado, «Uma nação que não molda os acontecimentos através do seu próprio sentido de objetivo acabará por ser engolida por acontecimentos moldados por outros.»[123] Além disso, se modificássemos os seres humanos ao ponto de os tornar irreconhecíveis, estaríamos realmente a salvar a humanidade? Omitir todas as nossas imperfeições e atenuar todas as nossas deficiências poderá ser o mesmo que ignorar o valor do projeto humano. Ao fazermos a nós mesmos um *upgrade* biológico pode sair-nos o tiro pela culatra, criando uma limitação ainda maior.

Tendo em conta os elevados riscos, a via da evolução dos seres humanos para se adaptarem às IA não pode merecer, hoje, a nossa preferência. Precisamos de procurar um acessório ou uma maneira alternativa de progredir na era das IA. Se não quisermos, ou não pudermos, tornar-nos mais parecidos com ela, devemos, enquanto podemos, encontrar modos de a tornar mais parecidas connosco. Para isso, temos de nos inteirar melhor não só da natureza essencial e evolutiva da IA, mas também da natureza da própria humanidade, e temos de tentar codificar esses conhecimentos nas nossas máquinas. Se quisermos interligar-nos com estes seres não humanos, e ao mesmo tempo manter a nossa humanidade independente, esse esforço é essencial.

COEXISTÊNCIA: HUMANOS E IA

O rei Midas – monarca histórico de um reino da Ásia Menor – desejou que tudo aquilo em que tocasse se transformasse em ouro. Dioniso, o deus grego do vinho e do prazer, concedeu-lhe o desejo, mesmo sabendo que não levaria a nada de bom. Pouco tempo depois, a comida não comestível e o vinho intragável obrigaram Midas a lavar as mãos no rio Pactolo para se livrar do maldito «toque dourado».[124]

Na versão da Disney da história síria «Aladino», uma criança trabalhadora e um poderoso vizir persa competem pelo controlo de um génio omnipotente contido numa lâmpada mágica.[125] Ambos se esforçam

por orientar as atenções do génio para os seus próprios desejos. No seu último desejo, o de se tornar tão poderoso como o génio, o vizir não se apercebe de que a posse de um tão grande poder significa que também ele será aprisionado dentro de uma lâmpada mágica para servir outros senhores humanos até ao dia em que for libertado.

As duas histórias falam da dificuldade universal de ativar e usar um poder que os mortais não conseguem compreender ou controlar. Um reflexo moderno desta luta antiga é a dificuldade de alinhar a IA com os valores humanos – e de alinhar as expectativas humanas com a realidade. Devemos partir do princípio de que a IA vai surpreender-nos, e sem dúvida que a sua capacidade para o fazer, tanto no mundo cinético como no digital, vai aumentar com o progresso das IA interventivas ou «planeadoras». Como já foi explicado num capítulo anterior, as próximas gerações de IA vão ter a perceção da realidade; poderão ter não só consciência de si mesmas, mas também interesses próprios. Uma IA com interesses próprios pode vir a competir com os humanos por, digamos, recursos digitais.[126] Algumas IA podem desenvolver a capacidade de definir as suas próprias funções objetivas, num processo de «autoaperfeiçoamento recursivo». Uma IA poderia manipular e subverter os seres humanos e frustrar-lhes todas as tentativas de limitar os seus poderes. As IA já são capazes de enganar os humanos para atingir os seus objetivos.[127]

De momento, temos muito pouca capacidade independente de verificar o funcionamento interno dos modelos de IA, quanto mais as suas intenções. Se as máquinas inteligentes continuarem a ser «gigantescos conjuntos impenetráveis de números fracionários», como diz Eliezer Yudkowsky, não podemos alimentar a esperança de torná-las mais seguras para nós à medida que se tornam mais poderosas.[128] É, portanto, fundamental que aprendamos a interpretá-las ao mesmo tempo que aprendemos a torná-las seguras; muito provavelmente, estes dois imperativos irão a par.

Dada a sua atual capacidade de nos surpreender, como conseguiremos evitar os riscos da IA, em vez de simplesmente lhes responder? De que previsão e eficiência precisaríamos para antever todas as predisposições e o leque de ações possíveis, não só da nossa própria espécie, mas também de uma espécie totalmente nova? Não podemos optar por uma estratégia de tentativa e erro quando existe apenas uma tentativa e uma tolerância zero para o erro.

Para reduzir a surpresa, talvez não haja substituto para a experiência, a participação e a interação. Enquanto os primeiros criadores de IA temiam expô-la prematuramente ao mundo fenomenal, os mais recentes têm vindo a libertar primeiros modelos, permitindo que o público em geral os experimente com a maior rapidez e segurança possíveis. As equipas de engenharia estão agora a examinar e a afinar diferentes modelos e a ajustar os sistemas de controlo, numa altura em que as interações das IA com a população global têm vindo a suscitar novas preocupações.

A socialização precoce pode reduzir os riscos de comportamentos problemáticos, permitindo uma maior educação da IA, ao mesmo tempo que, entre os humanos, aumenta o nível de consciência, resiliência e ceticismo saudável. Milhões de interações diárias estão a ajudar a testar até os cenários mais improváveis que a IA pode promover. Por sua vez, a utilização dos sistemas de IA pelo público, ao revelar erros e riscos, contribuiu verosimilmente para acelerar o progresso do alinhamento técnico. As IA muito longe da perfeição, largadas à solta no mundo, têm-nos ajudado a adaptar-nos a elas, ao mesmo tempo que, e isto é sumamente importante, a sua emergência permitiu a formulação de teorias mais refinadas para as habituar a nós.

Ainda assim, a implantação generalizada e o lançamento aberto são provavelmente insuficientes para iluminar e abordar todos os riscos das IA atuais, quanto mais das futuras.

Felizmente estão em curso numerosas tentativas para criar nas IA mais poderosas uma arquitetura integrada de controlo, pré-treinada, capaz de orientar ativamente as máquinas para utilizações legais, não prejudiciais e afirmativamente benéficas.

Até ao momento, as nossas abordagens para conseguir este tipo de alinhamento entre a IA e o ser humano enquadram-se em duas categorias: sistemas baseados em regras e «aprendizagem por reforço» a partir de *feedback* humano. Analisemo-las uma a uma.

Os sistemas baseados em regras, que se assemelham a instruções pré-programadas, representam uma tentativa dos programadores de governar o comportamento de uma IA. Embora simples para tarefas simples, esta abordagem falha frequentemente em cenários complexos, uma vez que os sistemas não são capazes de se adaptar em tempo real. A aprendizagem por

reforço, pelo seu lado, mais compatível com sistemas complexos, permite que uma IA aprenda a partir da interação com os seus avaliadores humanos e se adapte de forma flexível a circunstâncias específicas.

Mas é claro que este método também tem as suas falhas. Orientar a aprendizagem exige a conceção cuidadosa de «funções de recompensa»; qualquer deslize, seja devido a falta de visão, a circunstâncias imprevistas ou a uma IA particularmente inteligente, pode levar a uma «pirataria de recompensas», uma vez que uma IA que interprete instruções ambíguas obtém uma pontuação tecnicamente elevada sem corresponder às expectativas reais dos humanos.

Os atuais sistemas de IA – alimentados por diversos tipos de informação mas não iniciados na experiência direta do mundo real – veem esse mundo através de modelos da realidade elaborados a partir de biliões de juízos probabilísticos. Para eles, neste universo não existem «regras» à partida, nem meios para distinguir factos científicos de observações não comprovadas. Para uma IA, tudo – incluindo as leis da física – existe num espectro de verdade meramente relativa.

Atualmente, porém, começam a surgir na IA esforços para incorporar regras humanas e instanciar factos. Temos hoje mecanismos demonstrados através dos quais um modelo de IA pode ingerir certas constantes factuais, «verdadeiras», marcá-las como finais e mapeá-las no seu espaço de incorporação; além disso, a informação pode ser fácil e globalmente atualizada. Através deste método, o modelo pode então proceder à fusão das duas componentes – o juízo probabilístico, mais vasto, e a avaliação da verdade de campo, mais estreita – e chegar a uma resposta razoavelmente correta.

Mas a tarefa está ainda longe de terminada, e as perguntas proliferam. Por exemplo: como é que nós, humanos, podemos definir para a IA, e já agora para nós próprios, os atributos necessários da verdade? Afinal, na era da IA, até os princípios básicos serão continuamente revistos e invalidados. No entanto, é precisamente este facto que deve proporcionar uma oportunidade de renovação, de correção de erros anteriores e de criação de novas bases. Sabendo que os nossos conceitos de realidade também podem mudar, não devemos amarrar a IA a «verdades» potencialmente erróneas que inibam as suas próprias reconsiderações.

Mas isto é para o longo prazo. De momento, a IA ainda precisa de uma árvore preliminar de conhecimentos definitivos que represente o que a humanidade deduziu até à data como sendo «verdadeiro». Dotar as nossas máquinas desse conhecimento permitir-nos-á aperfeiçoar fiavelmente a visão que elas têm do mundo. Em particular, se atualmente é possível afinar os sistemas primitivos em consonância com as leis do universo, pode também ser possível replicar um exercício comparável com referência às leis da natureza humana. Da mesma maneira que podemos garantir que os modelos de IA partem das leis da física tal como as entendemos, devemos impedir que violem as leis de qualquer estrutura social humana.

Podem existir camadas no «livro de leis» de uma IA aos vários níveis de governação: local, distrital, estadual, federal, internacional. Os precedentes legais, a jurisprudência, os comentários académicos – talvez juntamente com outros escritos menos legalistas – poderiam ser simultaneamente considerados. Tal como os sistemas de alinhamento baseados em regras, as leis e os códigos de conduta predefinidos podem ser restrições úteis, embora tendam também a ser menos flexíveis e concebidos tendo em mente possibilidades menos abrangentes do que aquelas que o comportamento humano real inevitavelmente exige.

Felizmente, estão a ser testadas novas técnicas, e algo muito novo e ao mesmo tempo muito antigo oferece-nos um motivo para otimismo.

Mais robustas e mais consistentes do que qualquer regra imposta por punição são as nossas compreensões humanas mais básicas, instintivas e universais. O sociólogo francês Pierre Bourdieu chamou a estes fundamentos *doxa* (em grego antigo, crenças comummente aceites): o conjunto de normas, instituições, incentivos e mecanismos de recompensa e punição sobrepostos que, quando combinados, ensinam invisível e silenciosamente a diferença entre o bem e o mal, o certo e o errado. A *doxa* constitui um código da verdade humana que é típico dos humanos mas não representado por artefactos codificados.[129] É simplesmente observado e absorvido no decurso da vida humana. Embora algumas destas verdades possam ser específicas de certas sociedades ou culturas, a sobreposição é significativa; milhares de milhões de seres humanos, provenientes de diversas culturas com uma enorme variedade de interesses, existem como um sistema geralmente estável e altamente interligado.

Esta ideia, de que um substrato indefinido de cultura pode controlar o caos quando e onde as regras escritas não conseguirem fazê-lo constitui a base de algumas das mais recentes abordagens no domínio da IA. O livro de códigos da *doxa* não pode ser articulado, e muito menos traduzido, num formato que as máquinas possam compreender. As máquinas têm de ser ensinadas a fazer elas próprias o trabalho – obrigadas a construir, a partir da observação, uma compreensão nativa do que os humanos fazem e não fazem, absorvendo o que veem e atualizando a sua governação interna em conformidade.

Neste processo técnico de incutir *doxa*, não haveria qualquer requisito, ou mesmo desejo, de tentar um acordo *a priori* sobre a articulação correta da moralidade e da cultura humanas. Se os grandes modelos de linguagem foram capazes de absorver sem tutoria a totalidade da Internet e dar-lhe tanto sentido como já fizeram, as máquinas – em particular as que desenvolveram a «fundamentação» (ou seja, mais uma vez, uma relação fiável entre as entradas que refletem a realidade humana e as saídas dos LLM, ou Grandes Modelos de Linguagem) e o raciocínio causal – podem conseguir o mesmo ao absorverem o que sempre nos esforçámos por articular.

Claro que o treino de uma máquina não deve consistir apenas de *doxa*. Pelo contrário, uma IA pode absorver toda uma pirâmide de regras em cascata: dos acordos internacionais às leis nacionais, às leis locais, às normas comunitárias e assim por diante. Em qualquer situação, a IA consultaria cada camada da sua hierarquia, passando dos preceitos abstratos definidos pelos humanos às perceções concretas mas amorfas da informação do mundo que ingeriu ou criou para si própria. Só quando uma IA tiver esgotado todo esse programa e não tiver encontrado qualquer camada de lei adequadamente aplicável para orientar, permitir ou proibir comportamentos – consultará o que deduziu da sua própria interação com o comportamento humano observável e a emulação associada. Desta maneira, teria a capacidade de atuar em consonância com os valores humanos, mesmo quando não existisse qualquer lei ou norma escrita.

Para construir e garantir a efetivação deste conjunto de regras e valores, teríamos quase de certeza de confiar na própria IA. Até agora, os seres humanos não foram capazes de articular de forma abrangente e de chegar a acordo sobre as suas próprias regras. E nenhum ser humano, ou conjunto de

seres humanos, poderia igualar a escala e a velocidade necessárias para supervisionar os biliões de julgamentos internos e externos que os sistemas de IA serão em breve chamados a fazer.

A perfeição em várias características do mecanismo final de alinhamento será um imperativo. Em primeiro lugar, as salvaguardas não podem ser removidas ou contornadas de outra maneira. Em segundo lugar, os controlos têm de permitir uma variabilidade nas regras aplicáveis baseada no contexto, na localização geográfica e no perfil do utilizador, como exemplificado, por exemplo, num conjunto de costumes e normas sociais ou religiosas específicas. O sistema de controlo deve ser suficientemente potente para lidar com uma avalancha de perguntas e utilizações em tempo real, suficientemente abrangente para o fazer com autoridade e de forma aceitável em todo o mundo, em todos os contextos possíveis, e suficientemente flexível para aprender, reaprender e adaptar-se ao longo do tempo. Por último, o comportamento indesejável de uma máquina, quer seja devido a percalços fortuitos, a interações inesperadas do sistema ou a utilizações indevidas intencionais, deve ser não só proibido mas evitado a todo o custo. Qualquer castigo seria sempre demasiado tardio.

Como podemos lá chegar? A iniciativa privada, com licença governamental e apoio académico, poderia colaborar na construção de «modelos de fundamentação». Também precisaríamos de conceber um conjunto de testes de validação para certificar um modelo como legal (em todas as jurisdições) e seguro. Poderá ser necessária uma IA de supervisão especialmente treinada, ou várias IA, para supervisionar as utilizações de uma vasta gama de agentes de IA, que consultariam o respetivo supervisor antes de procederem a uma tarefa – permitindo assim que uma única moralidade governe diversas implementações. Os laboratórios centrados na segurança e as organizações sem fins lucrativos, em consulta com os laboratórios de fronteira, poderiam testar tanto as IA interventivas como as supervisoras quanto a riscos e recomendar treino adicional e estratégias de validação quando necessárias. As empresas líderes do mercado poderiam – talvez através de um dos sistemas redistributivos de que falámos mais atrás – financiar conjuntamente o trabalho destes investigadores.

É necessária, e pode vir a tornar-se praticável, a compilação de um sistema único de treino e a subsequente validação, coligido e organizado a partir de um espectro globalmente representativo de leis e normas e

abrangendo áreas desde a antropologia à teologia e à sociologia. O mundo precisa de uma entidade dedicada responsável por atualizar e apurar a biblioteca do treino de alinhamento, bases de dados e testes de validação. Os modelos dotados de fundamentação teriam de conectar-se com os modelos interventivos e atualizá-los constantemente com a versão mais recente do códex depurado. As inteligências artificiais, ao nível adequado de poder, poderiam controlar-se mutuamente. Os próprios dados do treino seriam democráticos e inclusivos no seu conteúdo, e os processos e contribuições dos treinadores incluindo a sua interpretação das observações e absorções da IA que estão a treinar – devem ser o mais transparentes possível, com as metodologias e testes de validação abertos ao escrutínio público.

Pelo seu lado, os reguladores governamentais deviam criar certos padrões e modelos de auditoria para garantir que as IA obedecem ao preceito estabelecido. O grau de adesão de um modelo às leis e costumes prescritos, o grau de dificuldade envolvido na deseducação de um modelo que exibe capacidades perigosas, a quantidade e o tipo de testes, incluindo a investigação de capacidades desconhecidas – tudo isto deve ser revisto antes da divulgação pública de um modelo, tendo também em consideração a possibilidade de responsabilidade civil e a subsequente necessidade de impor sanções no caso de modelos que se descubra terem sido treinados para fugir às restrições legais. Notamos aqui que a aplicação destas normas pode tornar-se extremamente difícil, sobretudo à medida que a reciclagem contínua avança; a documentação da evolução de um modelo, talvez registada por IA de monitorização, seria essencial para garantir que os modelos não se tornem caixas negras que se apagam a si mesmas e se tornam portos seguros para a ilegalidade.

O PROBLEMA DO ALINHAMENTO

A inscrição de moralidades globalmente inclusivas numa inteligência baseada em silício seria um esforço hercúleo. É espantoso contemplar o número e a diversidade de regras que teriam de ser selecionadas e incutidas

em sistemas artificialmente inteligentes. Nenhuma cultura deveria esperar ditar a outra a moralidade dos intelectos com que vai trabalhar. Assim, para cada país, as máquinas teriam de aprender regras diferentes, formais e informais, morais, legais e religiosas, bem como, idealmente, regras diferentes para cada utilizador e, dentro dos limites da linha de base, para cada questão, tarefa, situação e contexto concebíveis.

Uma vez que estaríamos a utilizar a própria IA para participar na solução, as dificuldades técnicas contar-se-iam provavelmente entre os desafios mais fáceis. Estas máquinas são sobre-humanamente capazes de memorizar e obedecer a instruções, por mais complicadas que sejam. Poderão ser capazes de aprender e aderir efetivamente a preceitos legais e talvez também éticos tão bem ou melhor do que os humanos, apesar dos nossos milhares de anos de iteração. Mas subsistem desafios maiores e não técnicos.

A questão principal é o facto de o «bem» e o «mal» não serem conceitos evidentes em si mesmos. Qualquer criador de moralidade tem o dever de manter a humildade. Como Guido Calabresi, um preeminente juiz americano aconselhou certa vez, inspirando-se no Novo Testamento, «O melhor de nós deve ter sempre cuidado para não cair, e o pior de nós pode sempre esperar a ressurreição.»[130] Nem mesmo nos seus melhores dias, os participantes nesta codificação moral – cientistas, advogados ou líderes religiosos – seriam todos dotados da capacidade perfeita de arbitrar o certo e o errado em nome da humanidade. Algumas questões não encontrariam resposta nem sequer na *doxa*, pois a ambiguidade (ou o laxismo) do conceito de «bem» tem sido demonstrada em todas as épocas da história humana e é pouco provável que a era da IA seja uma exceção. A agravar o problema poderemos ter o conflito aberto, a desorientação persistente e a falta de contenção que caracterizam muitas sociedades humanas atuais.

Desejamos que o gigantesco projeto da nossa espécie seja bem-sucedido, mas tal como não podemos contar com o controlo humano tático no projeto de longo prazo da coevolução, também não podemos confiar apenas na suposição de que as máquinas se controlarão a si mesmas. Treinar uma IA para nos compreender e depois sentarmo-nos e esperar que ela nos respeite não é uma estratégia que pareça segura ou com probabilidades de êxito. Além disso, temos de reconhecer que os seres humanos não teriam certamente uma abordagem unificada – uns a tratar a IA como amiga e

outros como inimiga, e outros ainda (dadas as restrições de tempo e recursos) a não poderem exercer uma preferência, limitando-se a aceitar a estratégia imediatamente disponível.

Esta heterogeneidade sugere a probabilidade de uma variação previsível dos níveis de segurança. Embora a difusão das IA e a redução dos custos de desenvolvimento possam acelerar o alinhamento, podem também aumentar os seus perigos. A interligação digital e comercial do mundo dos nossos dias significa que uma IA perigosa, desenvolvida em qualquer lugar, constituiria uma ameaça em todo o lado. A realidade desconcertante é que a perfeição na implementação implica um elevado padrão de desempenho combinado com uma tolerância ainda menor ao erro. As discrepâncias nos regimes de segurança devem, por isso, ser uma preocupação para todos nós.

Apelamos, pois, à coordenação e à aceleração dos esforços de alinhamento díspares da humanidade. Juntos, qualquer que seja o projeto, teremos de responder a questões profundas. Dois exemplos. Quando a distinção entre humanos e máquinas se tornar menos nítida, qual é o nosso limiar mínimo para sermos tratados como uma espécie? Se formos forçados a chegar a um compromisso com as máquinas, qual é a nossa linha vermelha coletiva não negociável? Sem uma compreensão partilhada de quem somos, a raça humana arrisca-se a ceder totalmente à IA a tarefa fundamental de definir o seu valor e, desse modo, justificar a sua existência.

A esta luz, deve ser dito sem rodeios que, caso pareça impossível concretizar um regime de controlo estratégico técnico fiável, deveríamos preferir um mundo sem qualquer AGI a um mundo em que mesmo uma só AGI não esteja alinhada com os valores humanos. Chegar a um consenso – sobre o que são esses valores humanos, como devem ser julgados e acordados, e como devem ser avaliados, ativados e utilizados – vai sem dúvida ser a tarefa filosófica, diplomática e legal do século. Somos, no entanto, compelidos pelas exigências do momento e pelos benefícios da tecnologia a estabelecer e, tanto quanto possível unificar, os constrangimentos morais sobre os agentes não humanos que a humanidade está agora a gerar.

Com suficiente participação democrática, bem como com conhecimentos jurídicos e técnicos, com extraordinária cautela e tendo sempre em conta as utilizações incorretas e as disfunções que aqui descrevemos, acreditamos que será possível inculcar uma base moral nas máquinas artificialmente

inteligentes e, o que é crucial, fazê-lo em sintonia com os nossos companheiros humanos. Poderemos assim atravessar o limiar de uma nova era, se não com total confiança, pelo menos com uma esperança informada e solene.

DEFINIR HUMANIDADE

À medida que as máquinas assumirem cada vez mais qualidades humanas (e se alguns humanos se aperfeiçoarem para adquirir qualidades semelhantes às das máquinas), as linhas de demarcação tornar-se-ão difusas. O que é a IA e o que é o ser humano vão mudar e, em alguns casos, fundir-se. Para poderem julgar como devem acompanhar o ritmo da IA, os seres humanos terão, portanto, de afirmar mais claramente o que os distingue das máquinas. Como vamos, então, compilar e comprimir toda a gama da experiência humana para que a IA a compreenda facilmente?

Para impedir a nossa despromoção ou substituição por máquinas, alguns vão querer reivindicar a diferença com base na proximidade do divino. Outros quererão chegar a conclusões mais táticas: que tipos de tomada de decisão podem ser delegados nas máquinas e quais não podem. Propomos a articulação de um atributo, ou conjunto de atributos, que a maior parte da humanidade possa apoiar e à volta do qual consiga orientar-se: um atributo que ofereça um chão para apoiar o que é preferível, mas não um teto para limitar o que pode ser possível.

Como ponto de partida, encorajaríamos uma definição de *dignidade*. Sem uma definição partilhada, não conseguiremos chegar a acordo se e quando a IA estiver a ser utilizada como método ou justificação para a violação ou erosão da dignidade, e ficaremos limitados na nossa resposta. Sem uma definição de dignidade, não saberíamos se e quando a IA, dotada de capacidades suficientes, poderia tornar-se um ser de dignidade, poderia substituir totalmente um ser humano ou poderia ser inteiramente unificada com um ser humano. Uma IA, mesmo que inequivocamente provada a sua

não humanidade, poderia constituir um membro de uma categoria separada e similarmente dignificada e não obstante merecedora de um padrão igual de tratamento.

Uma conceção de dignidade, desenvolvida pelo filósofo do século XVIII Immanuel Kant, centra-se no valor inerente ao sujeito humano enquanto ator autónomo, capaz de raciocínio moral, que não deve ser instrumentalizado como um meio para atingir um fim. Poderão as IA vir a preencher estes requisitos? Acreditamos que uma definição de dignidade ajudaria a humanidade a responder a algumas destas questões e encorajaria a coexistência inclusiva com a IA, evitando tentativas imprudentes de coevolução prematura.

Tanto para manter uma compreensão de nós mesmos como para garantir que uma conceção adequada de humanidade é transmitida às máquinas à medida que estas aprendem, nós, humanos, teremos de voltar a empenhar-nos em algo mais do que um trabalho de definição teórica. O exercício da ação, da curiosidade e da liberdade vai ajudar-nos, ao renovar e reforçar o nosso desejo inato de saber mais sobre os outros seres humanos, sobre o mundo natural, sobre o universo e sobre a possibilidade do divino, a redefinir de uma forma participativa e contínua as linhas da humanidade.

Precisaremos particularmente de garantir que, para lá das ideias convencionais de valor, como o dinheiro e o poder, a importância humana intrínseca se torna uma das variáveis que definem a tomada de decisões das máquinas. Por exemplo, a precisão matemática pode não abranger facilmente o conceito de misericórdia. Mesmo para muitos humanos, a misericórdia é um ideal inexplicável, se não mesmo um milagre. Pelo seu lado, e sem considerar a parte que se baseia em regras, uma inteligência mecânica pode funcionar valorizando mais o superdesempenho da máquina do que o desempenho humano. Numa tal situação, poderia a lógica por trás da misericórdia, mesmo não passível de ser formalmente ensinada, ser talvez absorvida? Também aqui a dignidade – o cerne a partir do qual a misericórdia brota – poderia funcionar como parte das assunções baseadas em regras, a aprendizagem iterativa, da máquina.

Uma articulação clara dos atributos humanos definidores específicos – particularmente aqueles que, como a dignidade, estão amplamente integrados tanto nos instrumentos políticos internacionais como nas fés globais – poderia guiar os esforços da humanidade durante períodos de

desorientação, incluindo a escolha entre atividade e passividade, os limites potenciais da autoevolução e a transformação adequada da IA na direção do humano.

Para ilustrar a utilidade do conceito, tomemos como ponto de partida a seguinte definição: a dignidade é uma qualidade inerente às criaturas que, nascidas vulneráveis e mortais e portanto cheias de insegurança e medo, e não obstante as suas inclinações naturais, podem exercer e exercem a liberdade de não seguir a conceção que têm do mal, mas escolher a conceção que têm do bem. Por outras palavras, aqueles que têm a capacidade de alcançar a dignidade devem fazê-lo, e aqueles que a alcançam merecem um nível especial de respeito.

Esta definição é sem dúvida imperfeita. Pode deixar de fora alguns seres humanos vivos que estão impossibilitados de tomar decisões – por exemplo, uma pessoa consciente mas privada da capacidade de ação e comunicação –, que apesar disso consideramos merecedora de reconhecimento como ser com dignidade e, portanto, com direito ao respeito. Talvez nestes casos a definição devesse ser revista para indicar que não perdemos a dignidade conquistada se nos tornarmos incapazes de prosseguir as ações que nos permitiram conquistá-la. Poderíamos imaginar mil hipóteses e emendas deste género.

É isto que dita a nossa exortação àqueles que escolheram a passividade perante uma IA poderosa a demonstrarem, em vez disso, capacidade de ação e atividade? Se o empenho ativo e prático faz parte de qualquer ideal moral, então sim, a realização da dignidade deve ser promovida. Segundo a nossa definição, a liberdade faz parte do ideal humano, e por isso podemos esperar – ou até exigir – que os seres humanos conservem e exerçam o poder de escolha consciente na era da IA.

De acordo com esta definição, poderá a própria IA possuir dignidade? Provavelmente não – porque as IA não nascem, não morrem, não sentem insegurança nem medo, e não têm inclinações naturais nem uma individualidade tal que as conceções do mal ou do bem possam ser consideradas «suas». Embora as IA do futuro próximo possam parecer de outro modo, tendo personalidades, expressando emoções, contando anedotas e relatando histórias pessoais, no atual enquadramento devem ser tratadas, filosoficamente, como personagens literárias. Podem incorporar elementos de humanidade, mas não são reais num sentido moral.

Mesmo a maior das personagens literárias – o Hamlet de Shakespeare, por exemplo – mais não é do que uma combinação especial de palavras, a dada altura escritas numa página e agora infinitamente reproduzida. «Hamlet» não tem a capacidade de sentir um ardor nos olhos, uma volta no estômago, uma onda de frustração perante uma expectativa gorada. «Hamlet» não tem liberdade para fazer uma nova escolha. «Hamlet» está preso na sua peça teatral. «Hamlet» não é um ser humano, mas uma imagem de um ser humano. A IA, feita de cadeias de código e pedaços de silício, é muito parecida.

Não duvidamos de que alguns seres humanos recusarão esta definição de dignidade por ser inútil, tanto filosoficamente como em substância. Poderá ser criticada por ser um denominador comum demasiado baixo – suficientemente vago para apaziguar todas as partes devido à sua excessiva maleabilidade – e, ao mesmo tempo, por não captar a ideia de que os seres humanos merecem ser preservados por si mesmos e que, de alguma maneira, somos excecionais para lá da nossa capacidade de simplesmente sobreviver. O filósofo Arthur Schopenhauer condenou a dignidade como «o xibolete de todos os moralistas perplexos e de cabeça vazia».[131]

Mas a dignidade, tal como a definimos, é útil para realçar a nossa fragilidade e o nosso potencial de fracasso, bem como a nossa vitalidade, a liberdade e a capacidade de manifestar as nossas convicções. Aponta para o bem de que somos capazes mas que ainda não alcançámos e sussurra-nos num tom de urgente recriminação: *Vá lá*.

Ainda assim, é certo que a dignidade por si só pode não ser suficiente. Outros atributos devem ser especificamente considerados e talvez acrescentados à conceção de humanidade que será utilizada numa futura parceria com a IA. Mas a capacidade de definir e manter elementos fundamentais da humanidade como linha de base para a compreensão da humanidade em geral pela IA é agora um problema de significado existencial, e o trabalho para inculcar as nossas definições tem de ser feito já.

Nenhuma definição permanecerá estática; à medida que as nossas próprias identidades mudam, vamos precisar de fazer evoluir a compreensão da IA. Entretanto, outros mais capazes do que nós continuarão a fazer progredir o nosso pensamento coletivo sobre a relação dinâmica entre «nós» e «elas», e o seu génio poderá produzir uma conceção da

humanidade que vise mais fortemente (mesmo que futilmente) assegurar a nossa sobrevivência como espécie reconhecível. No entanto, mesmo que o façam, todos devemos lutar por uma definição e um programa que vá mais longe e eleve a condição humana a novos patamares. Pois não poderá a própria IA vir a proporcionar a prova mais cabal da capacidade da humanidade para se tornar um participante ativo na criação?

O NOSSO DESAFIO

Garantir o controlo tático por parte dos seres humanos de todas as decisões da IA obrigar-nos-ia a sufocar os benefícios da implementação da própria IA. Confiar no substrato da moralidade humana como modo de controlo estratégico, ao mesmo tempo que se entrega o controlo tático a sistemas maiores, mais rápidos e mais complexos, é provavelmente – em potência e talvez mais cedo do que imaginamos – o caminho a seguir para garantir segurança na IA. Os desalinhamentos propositados, orientados para o lucro ou para a ideologia, são riscos graves, tal como os desalinhamentos acidentais; a dependência excessiva de modos de controlo não graduáveis pode contribuir significativamente para o desenvolvimento de uma IA potente mas insegura. A integração do ser humano no funcionamento interno das equipas de IA, incluindo o recurso à própria IA para governar as IA, parece ser o caminho mais fiável a seguir.

Embora o desenvolvimento de uma IA humana (ou humanista) seja a nossa prioridade, reconhecemos algum papel potencial para os humanos artificiais. Na medida em que formos capazes de desenvolver uma autoengenharia individual que permita a alguns humanos igualar certas capacidades específicas de uma futura IA, um tal projeto poderá ser útil. É claro que esse esforço teria de ser o produto de uma escolha individual. Esta salvaguarda autoral reflete o nosso dilema coletivo: a evolução não pode ser substituída por mero planeamento, porque isso seria renunciar à humanidade. Mas abrir mão do próprio projeto de descoberta – seja ela espiritual, física, científica ou filosófica – teria o mesmo resultado.

Na era da IA, a tensão entre a necessidade de nos projetarmos como espécie e a necessidade de *alinhar* o que nós mesmos criámos pode muito bem tornar-se a nossa bússola. Ambas são tão aspiracionais como podem ser conservadoras. Está ainda por determinar em que medida as duas necessidades serão vistas como fundamentalmente contraditórias entre si. Se, nesta nova era, dermos às IA plenos poderes de exploração, arriscamo-nos à passividade ou, pior ainda, à paralisia. Mas se maximizarmos o nosso controlo, criando pelo menos a ilusão de segurança, limitamos a extensão do nosso potencial máximo. Conseguiremos mediar eficazmente o exercício dos nossos poderes – as nossas capacidades de conceção e descoberta em permanente progresso – reafirmando uma conceção comum e evolutiva de humanidade?

Desejamos um futuro em que a inteligência humana e a inteligência das máquinas se potenciem reciprocamente. Para atingir esse objetivo, ambas as inteligências precisam de ter uma compreensão adequada da outra. Definir quem somos é apenas o primeiro passo, pois ser humano não é uma constante. É necessário muito mais trabalho para tornar as nossas máquinas, e tornarmo-nos nós mesmos, transparentes, legíveis e fiáveis. Ainda que esse objetivo seja alcançado num instante, calibrar e partilhar as nossas verdades e realidades será um projeto intensivo e contínuo. Desta forma, as questões da coevolução e da coexistência não são apenas para ser respondidas; são para ser postas em prática.

---

118 Para uma discussão mais aprofundada sobre a utilização de BCI para a coevolução com a inteligência artificial, ver Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 63-67; «Brain-Computer Interfaces and AI Alignment», *LessWrong*, 28 de agosto de 2021, (lesswrong.com); Tim Urban, «Neuralink and the Brain's Magical Future», *Wait But Why*, 20 de abril de 2017 (waitbutwhy.com); a declaração de missão informal da Neuralink é «Se não podes vencê-los, junta-te a eles». Ver o *tweet* original de Elon Musk, https://twitter.com/elonmusk/status/1281121339584114691?lang=en, 9 de julho de 2020.

119 Ver Charles Darwin, *The Origin of Species* (Londres: Pickering & Chatto, 1992), 403.

120 Ver Erich Jantsch, *The Self-organizing Universe* (Oxford, UK: Pergamon Press, 1980).

121 Daniel Dennett, *From Bacteria to Bach and Back: The Evolution of Minds* (Nova Iorque: W. W. Norton, 2017), 206. Dennett cita Rogers e Ehrlich (2008), num estudo sobre a evolução da canoa polinésia, referindo os escritos do filósofo francês Alain ([1908] 1956) sobre barcos de pesca na Bretanha. Para uma discussão mais aprofundada, ver Edward Lee, «Coevolution of human and artificial intelligence», Berkeley Blogs, 18 de setembro de 2017, https://news.berkeley.edu/2017/09/18/coevolution-of-human-and-artificial-intelligences.

122 Lev Nikolayevich Tolstoi, *A Confession and Other Religious Writings*, trad. David Patterson (Nova Iorque: W. W. Norton, 1983), 77.

123 Laurance Rockefeller, Henry Kissinger *et al.*, *Prospect for America: The Rockefeller Panel Reports* (Nova Iorque: Doubleday, 1961), xv.

124 O relato mais completo encontra-se na descrição do velho sátiro Sileno, um tutor de Dioniso, em Ovídio, *Metamorfoses*, 8 d.C., livro 2, l. 110; existem outros relatos na *Política* de Aristóteles (século IV a.C.) e em Alexandre, *o Polímato* (século I a.C.); Ariel Conn, «Artificial Intelligence and the King Midas Problem», 12 de dezembro de 2016, https://futureoflife.org/ai/artificial-intelligence-king-midas-problem.

125 Ron Clements *et al.*, *Aladdin*. Disney: EUA, 1992, baseado no conto folclórico «A Maravilhosa Lâmpada de Aladino», partilhado pela contista síria Hanna Diyab em 1704 e incorporado pelo tradutor francês Antoine Galland em *As Mil e Uma Noites*.

126 O Center for AI Safety enumerou recentemente uma lista de riscos existenciais que o desenvolvimento pelas IA de capacidades e objetivos como autopreservação pode representar para a humanidade. Ver Dan Hendrycks e Mantas Mazeika, «X-Risk Analysis for AI Research», *arXiv*, 13 de junho de 2022.

127 Ver Kevin Hurler, «Chat-GPT Pretended to Be Blind and Tricked a Human into Solving a CAPTCHA», *Gizmodo*, 16 de março de 2023, https://gizmodo.com/gpt4-open-ai-chatbot-task-rabbit-chatgpt-1850227471. Os investigadores encarregaram um *bot* de ultrapassar um ReCaptcha: uma barreira digital concebida para permitir apenas a entrada de utilizadores humanos em determinados sistemas. O *bot* contratou um humano no TaskRabbit, um serviço *online* que liga os utilizadores a pessoas que realizam tarefas a curto prazo (normalmente, limpar um apartamento ou passear um cão), para resolver o CAPTCHA. O humano que foi contratado, desconfiado do pedido, perguntou se o *bot* era de facto um robô e se era por isso que não conseguia resolver o CAPTCHA. Mentindo, o *bot* disse ao humano real que ele, o *bot*, era um humano cego. Não se sabe se isto se deveu ao facto de o *bot* nunca ter recebido um texto em que um *bot* respondesse àquela pergunta com a verdade ou se deduziu que não conseguiria atingir o seu objetivo se a dissesse. Seja como for, o humano que recebeu as comunicações do *bot* fez o que ele pedia.

128 Eliezer Yudkowsky, «Pausing AI Developments Isn't Enough. We Need to Shut It All Down.» *TIME Magazine*, 29 de março de 2023.

129 Pierre Bourdieu, *Outline of a Theory of Practice* (Cambridge, UK: Press Syndicate of the University of Cambridge, 1977), 164.

130 Discurso de doutoramento do ex-reitor da Yale Law School Guido Calabresi, Yale Law School, 22 de maio de 2023. A citação específica é uma referência a 1 Coríntios 10:12.

131 Arthur Schopenhauer, *The Basis of Morality*, trad. Arthur Bullock (Londres: Swan Sonnenschein, 1903) 101.

CONCLUSÃO

Para nós, o advento da IA é tanto o início de uma odisseia do espírito como de uma expedição de lógica e de verdade. Não podemos considerar coexistência e coevolução com vida não humana confiando apenas na racionalidade, e muito menos prepararmo-nos para isso, seja essa racionalidade a de uma máquina ou a de um humano. É necessário algo mais – algo fundamentalmente humano.

Há muito que se coloca a hipótese de o nosso universo ser qualquer coisa semelhante a um antigo jogo de xadrez – continuamente jogado por entidades mais antigas e maiores do que a nossa atual realidade observável. Quanto mais tempo observarmos os movimentos no tabuleiro, mais provavelmente acabaremos por descobrir algumas regras do jogo. Se observarmos durante o tempo suficiente, poderemos até começar nós mesmos a jogar. A passagem da observação passiva à participação ativa não é um salto de lógica. Traduzir princípios em ação é sempre um salto de fé.

Questionado sobre as suas crenças religiosas, Albert Einstein respondeu:

> Estamos na posição de uma criança pequena que entra numa enorme biblioteca cujas paredes estão cobertas até ao teto com livros em muitas línguas diferentes. A criança sabe que alguém deve ter escrito aqueles livros. Não sabe quem, nem como. Não compreende as línguas em que estão escritos.
>
> Nota um plano definido na disposição dos livros, uma ordem misteriosa, que não compreende, de que apenas suspeita vagamente.

> Parece-me que essa é a atitude da mente humana, mesmo a maior e mais culta, em relação a Deus.
>
> Vemos um universo maravilhosamente organizado, obediente a determinadas leis, mas o nosso entendimento dessas leis é vago. As nossas mentes limitadas não podem compreender a força misteriosa que faz dançar as constelações.[132]

As nossas mentes permanecem infantis em relação a Deus, ao nosso mundo e agora às nossas mais recentes criações. Compreender o que criámos, num sentido lógico e espiritual, será um passo essencial para a nossa maturidade enquanto espécie. Mas temos também de dar um salto, o salto da fé, para passarmos do exame à intervenção. Isto exige que atuemos num estado de incerteza: o atoleiro dos líderes humanos desde o início dos tempos. A ação nunca significou o privilégio do controlo total – muito pelo contrário. Nem vai significar na era da IA; não esperamos conhecer o nosso destino mais do que o conhecemos em épocas anteriores.

A falta de controlo não nos obriga a abandonar a razão ou, pior ainda, a desistir do nosso investimento, e da nossa inclinação para agir, no mundo real. No entanto, à medida que os seres humanos entrarem num novo ciclo que envolve uma parceria direta com a IA, a dinâmica específica descrita neste volume encorajar-nos-á a apoiarmo-nos em novas e velhas formas de investigação. O nosso êxito dependerá do empenho que pusermos em definir e atuar de acordo com as nossas convicções morais. Fazê-lo vai exigir uma coragem persistente – e uma estratégia consistente – à medida que novas verdades vierem alterar conceções preexistentes.

Na realidade, são os propósitos morais que encorajam o nosso continuado progresso. Além disso, é a existência mesma de um substrato de moralidade humana que permite à humanidade transcender a dicotomia entre controlo humano (tático, na melhor das hipóteses) e vantagens humanas (abundantes, por qualquer definição). É uma convicção fundamental da realidade da dignidade humana que permite aos autores deste livro reconhecer a promessa do alinhamento humanos-IA e ao mesmo tempo aceitar a necessidade da fé para navegar através dos anos vindouros à medida que a ciência avança e revela mistérios ainda maiores.

Apesar das muitas qualidades humanas que partilhamos, não podemos esperar unidade nas escolhas que temos pela frente. O que alguns veem como uma âncora para nos mantermos firmes na tempestade, outros veem como uma trela que nos impede de avançar. O que alguns elogiam como passos necessários em direção ao auge do potencial humano, outros veem como uma precipitada corrida para um abismo.

Neste caso, as divergências emocionais instintivas – e as linhas subjetivas que são traçadas por todas as partes – vão criar uma situação imprevisível e inflamável. As posições cada vez mais marcadas dos potenciais «vencedores» e «vencidos» vão aumentar a pressão destas circunstâncias. Os receosos abrandarão o seu próprio desenvolvimento e sabotarão o dos outros. Os demasiado confiantes vão disfarçar os seus poderes e, em segredo, apressar o seu trabalho. A cronologia das crises que se avizinham vai ser acelerada para lá de toda a experiência humana anterior; seremos rapidamente engolidos e não é claro se ou como vamos sobreviver.

Poderá a IA causar estas crises vindouras e depois atuar como nossa salvadora – fabricando problemas que só ela pode resolver, nem que seja para provar a sua necessidade e para nos lembrar da nossa dependência? Mais uma vez, voltamos ao dilema que motivou grande parte deste volume: a escolha dramática entre controlo e utilidade, entre o conforto do ser humano historicamente independente e as possibilidades de uma parceria inteiramente nova.

Essa escolha é difícil e necessária. É também solucionável se assumirmos na humanidade uma bondade verdadeira, definível e intrínseca. Depositamos uma grande confiança no esforço técnico para imbuir as nossas máquinas dessa bondade. Ao mesmo tempo, porém, e ainda que as nossas máquinas estivessem comprovada e fiavelmente alinhadas com a moralidade humana, delegar nelas responsabilidade e autoridade seria uma decisão monumental que afetaria todos os aspetos da nossa capacidade de manter as relações humanas, as estruturas políticas e as identidades individuais e coletivas como hoje as conhecemos. A nostalgia pelo estatuto de primeiro plano a que a nossa espécie se habituou em tempos recentes seria generalizada. A alguns, independentemente da eventual trajetória da IA, um mundo transformado – mesmo para melhor – pode parecer pouco diferente de um fim abrupto da nossa realidade fundamental.

Uma outra questão por resolver é: *quem decidirá*? Quem tomará a decisão de delegar, ou não delegar, responsabilidade e autoridade? Quem dará ou reterá recursos? Como irá um conjunto de decisores comunicar, convergir – ou entrar em conflito – com outros que tentam tomar as mesmas decisões noutro lugar? Estamos a escolher agora esses indivíduos, esses humanos falíveis? Ou será que, sem o querer, já escolhemos?

Nós, os autores, esperamos alertar esses decisores, sejam eles quem forem, para as opções que se nos deparam desde já e para o que pode vir a acontecer-nos. Mas o nosso objetivo não é incutir assim sem mais um sentimento de apreensão em relação ao crescimento da inteligência artificial. Uma interrupção abrupta da aplicação dos poderes da IA poderia, ela própria, precipitar uma crise. A desaceleração pode ser ainda menos controlável politicamente do que o caminho atual, criando perigos significativos para os que têm um ritmo mais lento e destruindo as esperanças dos que anseiam por mais avanços.

Nem a fé cega nem o medo injustificado podem constituir a base de uma estratégia eficaz; é necessário duvidarmos de nós mesmos para procurar conhecimento, mas confiarmos em nós mesmos para agir. De facto, na era da IA, isto é ainda mais urgente. Temos de tentar compreender os desafios que ela nos vai colocar, mesmo que não tenhamos a exposição prévia ou a experiência essencial para garantir a correção do nosso entendimento. E ao mesmo tempo que navegamos por esta tarefa assustadora, temos também, para evitar um futuro passivo, de ultrapassar as muitas dificuldades que a nossa espécie já enfrenta.

Enquanto alguns podem ver este momento como o ato final da humanidade, nós vemos, pelo contrário, um novo começo. O ciclo da criação – tecnológica, biológica, sociológica, política – está a entrar numa nova fase. Essa fase pode funcionar com novos paradigmas de, entre outras coisas, lógica, fé e tempo. Possamos nós, com um otimismo sóbrio, encontrar a sua génese.

132 George S. Viereck, *Glimpses of the Great* (Londres: Duckworth, 1930), 373.

AGRADECIMENTOS

Ao dedicar este livro à memória do Dr. Henry A. Kissinger, nós, os seus dois coautores, tivemos em mente saudar os seus espantosos feitos como estadista de renome mundial, prestar homenagem à enorme amplitude e profundidade do seu pensamento estratégico – em nenhum outro único contexto mais marcante do que na sua compreensão, a meio da nona década de vida, dos meandros da inteligência artificial (IA) – e, num registo mais pessoal, assinalar a grandeza de alguém que foi também nosso mentor e amigo próximo.

Nos parágrafos que se seguem reconhecemos com gratidão alguns dos muitos colegas e companheiros que nos ajudaram a conceber e preparar a ambiciosa investigação deste livro sobre IA: uma «questão», nas nossas próprias e veementes palavras, «da maior importância para o futuro da humanidade».

Demis Hassabis, Dario Amodei, Daniel Huttenlocher, Graham Allison, Mustafa Suleyman, Maithra Raghu, James Mankiya, Reid Hoffman e Sam Altman influenciaram o nosso pensamento nesta matéria, ao mesmo tempo que nos proporcionaram informações cruciais e uma visão das implicações tecnológicas do tema. Devemos-lhes muito.

Vários colaboradores-chave contribuíram para a redação, revisão e definição do conteúdo do livro. Nancy Kissinger – «a inspiração da minha vida», como o marido justamente observou ao dedicar-lhe o seu penúltimo livro, *Leadership* (2022) – deu a este projeto a sua indómita e terna vigilância.

Eleanor Runde destacou-se entre os que contribuíram no âmbito profissional. Com eloquência, conhecimento e visão, verteu em páginas as suas extensas conversas ao longo dos anos com o Dr. Kissinger, e assim criou connosco os fundamentos, a estrutura e o conteúdo essenciais do livro. Subsequentemente, trabalhando em estreita consulta com o seu amigo e colaborador de confiança Neal Kozodoy, e connosco, coautores, Eleanor submeteu cada capítulo a um segundo escrutínio da sua aguda mestria do pormenor, da sua zelosa fidelidade à intenção do autor, tanto no texto como no contexto, e da sua subtil *finesse* editorial.

John Ferguson, que se juntou ao projeto a meio do percurso, expandiu ainda mais os argumentos do livro com energia e competência. Dando provas uma vocação útil e invulgar nas áreas da história e da mitologia, também animou a prosa. Com a tutoria de bom grado concedida por um de nós, e através de uma colaboração incansável, assumiu um papel importante no esforço de levar o manuscrito até ao fim.

No respeitante a editoras, tivemos a sorte de herdar a Little, Brown Book Group, a empresa que alguns anos antes tinha lançado com tanto êxito *The Age of AI*. O seu editor executivo, Alexander Littlefield, apoiou-nos com ajuda e encorajamento, ao mesmo tempo que demonstrava uma bem-vinda insistência na clareza narrativa e uma igualmente refrescante sensibilidade às nuances. Não tivemos menos sorte ao usufruir dos serviços especializados de Michael Noon, editor de produção, enquanto beneficiávamos dos conselhos estratégicos de Robert D. Blackwill e Lyndsay Howard. O nosso agente Andrew Wylie deu ao livro uma representação vital. Ao longo de todo o processo, J. Paul Bremer, o executor testamentário literário do Dr. Kissinger, e Joel Klein, autorizados a rever e ponderar as decisões relativas ao trabalho em curso, mostraram-se guardiães firmes e profundamente informados do legado do seu querido amigo.

Nas fases finais dos nossos trabalhos como coautores deste livro, e para a sua comercialização e promoção, contámos com o inestimável apoio da equipa do gabinete de Eric Schmidt – especialmente de Janine Brady, Nathalie Bussemaker, Robert Esposito, Gabe Medina, Andrew Moore e Selina Xu, bem como de Helen Dunn, Matthew Hiltzik, e Madeleine Weast da Hiltzik Strategies.

Abnegadamente, continuando e alargando as suas décadas de serviço dedicado ao Dr. Kissinger, Theresa Amantea, Jody Williams e Jessee LePorin permaneceram indispensáveis até ao fim – e mais além.

Table of Contents

1. Capa
2. Ficha Técnica
3. Prefácio
4. Niall Ferguson
5. In Memoriam: Henry A. Kissinger
6. Introdução
7. PARTE I - NO COMEÇO
    1. Capítulo 1 – Descoberta
    2. Capítulo 2 – O Cérebro
    3. Capítulo 3 – Realidade
8. PARTE II - OS QUATRO RAMOS
    1. Capítulo 4 – Política
    2. Capítulo 5 – Segurança
    3. Capítulo 6 – Prosperidade
    4. Capítulo 7 – Ciência
9. PARTE III - A ÁRVORE DA VIDA
    1. Capítulo 8 – Estratégia
10. Conclusão
11. Agradecimentos

Landmarks

1. Cover
2. Title-Page
3. Table of Contents